

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2003-030184

(43)Date of publication of application : 31.01.2003

(51)Int.Cl.

G06F 17/27

(21)Application number : 2001-217619

(71)Applicant : SONY CORP

(22)Date of filing : 18.07.2001

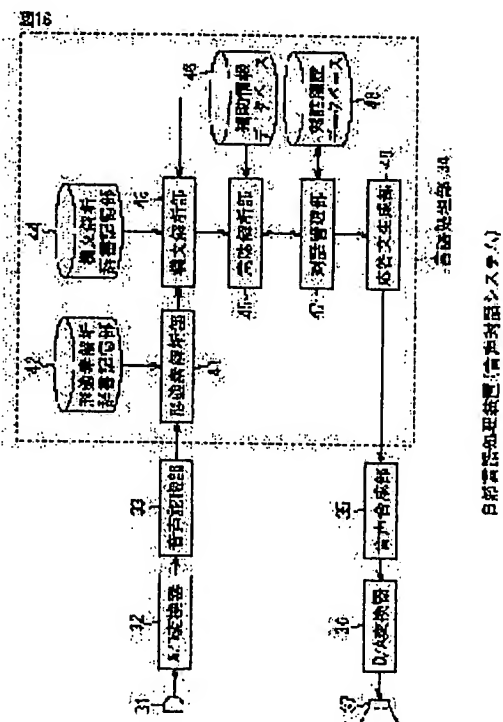
(72)Inventor : TAJIMA KAZUHIKO
YOKOTA SHIGEAKI
SHIMOMURA HIDEKI

(54) DEVICE/METHOD FOR PROCESSING NATURAL LANGUAGE, PROGRAM AND RECORDING MEDIUM

(57)Abstract:

PROBLEM TO BE SOLVED: To precisely understand the meaning of an inputted sentence by precisely performing syntax analysis and semantic analysis.

SOLUTION: A semantic analysis part 45 retrieves auxiliary information on a verb included in the inputted sentence from an auxiliary information database 46 storing auxiliary information generated by using a large quantity of corpus data consisting of subcategorization information and term structure information of verbs, and recognizes an anaphora-type attribute included in the inputted sentence on the basis of the auxiliary information concerning the verb included in the inputted sentence. Then, the semantic analysis part 45 determines an antecedent indicated by the coincident type and performs the semantic analysis of the inputted sentence by using the precedent.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

* NOTICES *

JPO and INPIT are not responsible for any damages caused by the use of this translation.

1. This document has been translated by computer. So the translation may not reflect the original precisely.
2. **** shows the word which can not be translated.
3. In the drawings, any words are not translated.

CLAIMS

[Claim(s)]

[Claim 1] The morphological analysis means which is natural-language-processing equipment which searches for the auxiliary information which assists the analysis of natural language from corpus data, and carries out morphological analysis of said corpus data, A basic sentence generation means to generate the basic sentence which is the unit made applicable [of a case frame] to generation from the morphological analysis result of said corpus data, An unnecessary lexical deletion means to delete an unnecessary vocabulary from said basic sentence to generation of a case frame, A case frame generation means to generate a case frame about the verb in the basic sentence from which said unnecessary vocabulary was deleted, Natural-language-processing equipment characterized by having an auxiliary information generation means to generate the subcategorization information and argument structure information on the verb, and to output as said auxiliary information, based on the case frame about the same verb.

[Claim 2] Said auxiliary information generation means is natural-language-processing equipment according to claim 1 characterized by for the verb generating the classification information showing whether it is what classified into any of an intransitive verb, a transitive verb, a **** verb, or the double object transitive verbs, and generating said subcategorization information based on said classification information based on the case frame about the same verb.

[Claim 3] the vocabulary which said unnecessary lexical deletion means becomes from an adverb, a noun, and "***", the vocabulary which consists of a noun, a particle, and "***", an adjective, and a noun -- "-- the natural-language-processing equipment according to claim 1 characterized by to delete the vocabulary which consists of the vocabulary which consists of ", the vocabulary which consists of a noun and a postposition, the part enclosed in the parenthesis or a part enclosed in the parenthesis, and "***" from said basic sentence.

[Claim 4] Said auxiliary information generation means is natural-language-processing equipment according to claim 1 characterized by generating said subcategorization information based on the case-marking particle of the case frame about the same verb.

[Claim 5] Said auxiliary information generation means is natural-language-processing equipment according to claim 1 characterized by generating said argument structure information based on the particles of all the case frames about the same verb.

[Claim 6] Said corpus data are natural-language-processing equipment according to claim 1 characterized by being Japanese data.

[Claim 7] The morphological analysis step which is the natural-language-processing approach of searching for the auxiliary information which assists the analysis of natural language from corpus data, and carries out morphological analysis of said corpus data, The basic sentence generation step which generates the basic sentence which is the unit made applicable [of a case frame] to generation from the morphological analysis result of said corpus data, The unnecessary lexical deletion step which deletes an unnecessary vocabulary from said basic sentence to generation of a case frame, The case frame generation step which generates a case frame about the verb in the basic sentence from which said

unnecessary vocabulary was deleted, The natural-language-processing approach characterized by having the auxiliary information generation step which generates the subcategorization information and argument structure information on the verb, and is outputted as said auxiliary information based on the case frame about the same verb.

[Claim 8] The morphological analysis step which is the program to which natural language processing which searches for the auxiliary information which assists the analysis of natural language from corpus data is made to carry out to a computer, and carries out morphological analysis of said corpus data, The basic sentence generation step which generates the basic sentence which is the unit made applicable [of a case frame] to generation from the morphological analysis result of said corpus data, The unnecessary lexical deletion step which deletes an unnecessary vocabulary from said basic sentence to generation of a case frame, The case frame generation step which generates a case frame about the verb in the basic sentence from which said unnecessary vocabulary was deleted, The program characterized by having the auxiliary information generation step which generates the subcategorization information and argument structure information on the verb, and is outputted as said auxiliary information based on the case frame about the same verb.

[Claim 9] The morphological analysis step which is the record medium with which the program to which natural language processing which searches for the auxiliary information which assists the analysis of natural language from corpus data is made to carry out to a computer is recorded, and carries out morphological analysis of said corpus data, The basic sentence generation step which generates the basic sentence which is the unit made applicable [of a case frame] to generation from the morphological analysis result of said corpus data, The unnecessary lexical deletion step which deletes an unnecessary vocabulary from said basic sentence to generation of a case frame, The case frame generation step which generates a case frame about the verb in the basic sentence from which said unnecessary vocabulary was deleted, The record medium characterized by recording the program equipped with the auxiliary information generation step which generates the subcategorization information and argument structure information on the verb, and is outputted as said auxiliary information based on the case frame about the same verb.

[Claim 10] An auxiliary information storage means by which are natural-language-processing equipment which carries out natural language processing of the input statement, and the auxiliary information which consists of the subcategorization information and argument structure information on verbal is memorized at least, A retrieval means to retrieve said auxiliary information about the verb contained in said input statement from said auxiliary information storage means, An attribute recognition means to recognize the attribute of a judgment means to judge whether an anaphor exists in said input statement, and the anaphor which exists in said input statement based on said auxiliary information about the verb contained in the input statement, Natural-language-processing equipment characterized by having an antecedent decision means to determine the antecedent to which said anaphor points, and an analysis means to perform syntax analysis or the semantic analysis of said input statement using the antecedent determined in said antecedent decision means, based on the attribute of said anaphor.

[Claim 11] Said judgment means is natural-language-processing equipment according to claim 10 characterized by judging whether an anaphor exists in said input statement based on the subcategorization information on the syntax-analysis result of said input statement, or said auxiliary information about the verb contained in said input statement.

[Claim 12] Said anaphor is natural-language-processing equipment according to claim 10 characterized by being a pronoun or a zero anaphor.

[Claim 13] It is natural-language-processing equipment according to claim 10 characterized by being dialogue equipment which has a dialog while memorizing dialogue hysteresis, and said antecedent decision means determining said antecedent by referring to said dialogue hysteresis.

[Claim 14] Based on a reply of a user [as opposed to / have further an inquiry means to ask the contents of said antecedent, to a user, and / said inquiry in said antecedent decision means], it is natural-language-processing equipment according to claim 10 characterized by determining said antecedent.

[Claim 15] It is the natural-language-processing approach which carries out natural language processing of the input statement. At least The retrieval step which retrieves said auxiliary information about the verb contained in said input statement from an auxiliary information storage means by which the auxiliary information which consists of the subcategorization information and argument structure information on verbal is memorized, The judgment step which judges whether an anaphor exists in said input statement, The attribute recognition step which recognizes the attribute of the anaphor which exists in said input statement based on said auxiliary information about the verb contained in the input statement, The natural-language-processing approach characterized by having the antecedent decision step which determines the antecedent to which said anaphor points, and the analysis step which performs syntax analysis or the semantic analysis of said input statement using the antecedent determined in said antecedent decision step based on the attribute of said anaphor.

[Claim 16] It is the program to which natural language processing which carries out natural language processing of the input statement is made to carry out to a computer. The retrieval step which retrieves said auxiliary information about the verb contained in said input statement from an auxiliary information storage means by which the auxiliary information which consists of the subcategorization information and argument structure information on verbal is memorized at least, The judgment step which judges whether an anaphor exists in said input statement, The attribute recognition step which recognizes the attribute of the anaphor which exists in said input statement based on said auxiliary information about the verb contained in the input statement, The program characterized by having the antecedent decision step which determines the antecedent to which said anaphor points, and the analysis step which performs syntax analysis or the semantic analysis of said input statement using the antecedent determined in said antecedent decision step based on the attribute of said anaphor.

[Claim 17] It is the record medium with which the program to which natural language processing which carries out natural language processing of the input statement is made to carry out to a computer is recorded. The retrieval step which retrieves said auxiliary information about the verb contained in said input statement from an auxiliary information storage means by which the auxiliary information which consists of the subcategorization information and argument structure information on verbal is memorized at least, The judgment step which judges whether an anaphor exists in said input statement, The attribute recognition step which recognizes the attribute of the anaphor which exists in said input statement based on said auxiliary information about the verb contained in the input statement, The antecedent decision step which determines the antecedent to which said anaphor points based on the attribute of said anaphor, The record medium characterized by recording the program equipped with the analysis step which performs syntax analysis or the semantic analysis of said input statement using the antecedent determined in said antecedent decision step.

[Translation done.]

* NOTICES *

JPO and INPIT are not responsible for any damages caused by the use of this translation.

- 1.This document has been translated by computer. So the translation may not reflect the original precisely.
- 2.**** shows the word which can not be translated.
- 3.In the drawings, any words are not translated.

DETAILED DESCRIPTION

[Detailed Description of the Invention]

[0001]

[Field of the Invention] This invention enables it to acquire the subcategorization information and argument structure information about a verb about a program and a record medium, further, determines the antecedent of an anaphor using the subcategorization information and argument structure information, and relates [natural-language-processing equipment and the natural-language-processing approach, and a list] to a program and a record medium at the natural-language-processing equipment and the natural-language-processing approach of enabling it to perform natural language processing, such as a high dialogue of precision, and a translation, and a list.

[0002]

[Description of the Prior Art] With conventional natural-language-processing equipment, morphological analysis of the inputted sentence (input statement) is carried out, further, syntax analysis and a semantic analysis are performed based on the morphological analysis result, and the semantic content of an input statement is understood. And when natural-language-processing equipment is dialogue equipment which performs a dialogue with a user, based on an understanding of the semantic content of an input statement, the response sentence to the input statement is generated and outputted.

[0003]

[Problem(s) to be Solved by the Invention] A subject called someone and a direct object called something in what was eaten lack having eaten in the input statement "whether it has already eaten" in the place. Therefore, about this input statement "has it already eaten?", unless it can determine that missing subject and direct object, it cannot be said that he understood that semantics to accuracy.

[0004] Here, according to the natural language theory of the publications in the Koichi Hashida "Global Document Annotation;GDA" Electrotechnical Laboratory, 1998 [for example, science 6 "generative grammar" Iwanami Shoten of Iwanami lecture-language, 1997,], etc., a thing like the pronoun which is called a zero anaphor (zero anaphora) and which it is [pronoun] in the location of an object and forms an anaphoric relation although not expressed exists. That is, in this natural language theory, when the noun phrase which should be in a certain location is missing, it is treated noting that a zero anaphor (zero anaphora) exists in that location.

[0005] In addition, a correspondence (anaphora) is a language phenomenon expressed by the group of alternative representation (anaphor), such as a pronoun and a demonstrative, and the object (antecedent) of those to which it points, and the anaphor which is not expressed is a zero anaphor.

[0006] In order to understand an above-mentioned input statement "has it already eaten?" to accuracy For example, now, if a zero anaphor is expressing pro, it will be set to syntax analysis. The verb "it eats" in an input statement "has it already eaten?" is classified on the basis of whether it is what needs what kind of constituent. based on the classification result, if an input statement "has it already eaten?" is "whether did also obtain pro (subject) and to have eaten pro (direct object)", it needs to analyze (analysis). Furthermore, when a zero anaphor (pro) exists, the antecedent to which the zero anaphor points needs to determine what it is concretely. It is necessary to specifically determine who ate, whether

it ate, and what it is about an input statement "has it already eaten?"

[0007] The intransitive verb which takes an agent (Agent) to a subject as a verbal classification here (intransitive), The **** verb (ergative) which takes an object (Theme) to a subject, the transitive verb which chooses a direct object (transitive), And there are four of the double object transitive verbs (ditransitive) which choose both a direct object and an indirect object, and it means classifying a verb into these intransitive verbs, a **** verb, a transitive verb, or the double object transitive verbs as classifying a verb. In addition, an above-mentioned verb "it eats" is a transitive verb.

[0008] However, in Japanese, since a subject and an object were omitted frequently, with conventional natural-language-processing equipment, analysis which took into consideration the verbal classification and the zero anaphor also in a surface or the depths at the time of syntax analysis was seldom performed.

[0009] Therefore, with conventional natural-language-processing equipment, since also judging the existence of the zero anaphor in an input statement and determining the antecedent when there is a zero anaphor further were also seldom performed, he was not able to understand semantics of an input statement to accuracy in many cases, without the ability performing high syntax analysis or the high semantic analysis of precision.

[0010] This invention is made in view of such a situation, makes possible high syntax analysis and the high semantic analysis of precision, and enables it to understand the semantics of an input statement to accuracy further by that cause.

[0011]

[Means for Solving the Problem] A basic sentence generation means by which the 1st natural-language-processing equipment of this invention generates the basic sentence which is the unit made applicable [of a case frame] to generation from the morphological analysis result of corpus data, An unnecessary lexical deletion means to delete an unnecessary vocabulary from a basic sentence to generation of a case frame, A case frame generation means to generate a case frame about the verb in the basic sentence from which the unnecessary vocabulary was deleted, Based on the case frame about the same verb, the subcategorization information and argument structure information on the verb are generated, and it is characterized by having an auxiliary information generation means to output as auxiliary information.

[0012] The basic sentence generation step to which the 1st natural-language-processing approach of this invention generates the basic sentence which is the unit made applicable [of a case frame] to generation from the morphological analysis result of corpus data, The unnecessary lexical deletion step which deletes an unnecessary vocabulary from a basic sentence to generation of a case frame, The case frame generation step which generates a case frame about the verb in the basic sentence from which the unnecessary vocabulary was deleted, Based on the case frame about the same verb, the subcategorization information and argument structure information on the verb are generated, and it is characterized by having the auxiliary information generation step outputted as auxiliary information.

[0013] The basic sentence generation step to which the 1st program of this invention generates the basic sentence which is the unit made applicable [of a case frame] to generation from the morphological analysis result of corpus data, The unnecessary lexical deletion step which deletes an unnecessary vocabulary from a basic sentence to generation of a case frame, The case frame generation step which generates a case frame about the verb in the basic sentence from which the unnecessary vocabulary was deleted, Based on the case frame about the same verb, the subcategorization information and argument structure information on the verb are generated, and it is characterized by having the auxiliary information generation step outputted as auxiliary information.

[0014] The basic sentence generation step to which the 1st record medium of this invention generates the basic sentence which is the unit made applicable [of a case frame] to generation from the morphological analysis result of corpus data, The unnecessary lexical deletion step which deletes an unnecessary vocabulary from a basic sentence to generation of a case frame, The case frame generation step which generates a case frame about the verb in the basic sentence from which the unnecessary vocabulary was deleted, Based on the case frame about the same verb, the subcategorization information and argument structure information on the verb are generated, and it is characterized by recording the

program equipped with the auxiliary information generation step outputted as auxiliary information.

[0015] A retrieval means by which the 2nd natural-language-processing equipment of this invention retrieves at least the auxiliary information about the verb contained in an input statement from an auxiliary information storage means by which the auxiliary information which consists of the subcategorization information and argument structure information on verbal is memorized, An attribute recognition means to recognize the attribute of a judgment means to judge whether an anaphor exists in an input statement, and the anaphor which exists in an input statement based on the auxiliary information about the verb contained in the input statement, Based on the attribute of an anaphor, it is characterized by having an antecedent decision means to determine the antecedent to which an anaphor points, and an analysis means to perform syntax analysis or the semantic analysis of an input statement using the antecedent determined in the antecedent decision means.

[0016] The retrieval step at which the 2nd natural-language-processing approach of this invention retrieves at least the auxiliary information about the verb contained in an input statement from an auxiliary information storage means by which the auxiliary information which consists of the subcategorization information and argument structure information on verbal is memorized, The attribute recognition step which recognizes the attribute of the judgment step which judges whether an anaphor exists in an input statement, and the anaphor which exists in an input statement based on the auxiliary information about the verb contained in the input statement, Based on the attribute of an anaphor, it is characterized by having the antecedent decision step which determines the antecedent to which an anaphor points, and the analysis step which performs syntax analysis or the semantic analysis of an input statement using the antecedent determined in the antecedent decision step.

[0017] The retrieval step at which the 2nd program of this invention retrieves at least the auxiliary information about the verb contained in an input statement from an auxiliary information storage means by which the auxiliary information which consists of the subcategorization information and argument structure information on verbal is memorized, The attribute recognition step which recognizes the attribute of the judgment step which judges whether an anaphor exists in an input statement, and the anaphor which exists in an input statement based on the auxiliary information about the verb contained in the input statement, Based on the attribute of an anaphor, it is characterized by having the antecedent decision step which determines the antecedent to which an anaphor points, and the analysis step which performs syntax analysis or the semantic analysis of an input statement using the antecedent determined in the antecedent decision step.

[0018] The retrieval step at which the 2nd record medium of this invention retrieves at least the auxiliary information about the verb contained in an input statement from an auxiliary information storage means by which the auxiliary information which consists of the subcategorization information and argument structure information on verbal is memorized, The attribute recognition step which recognizes the attribute of the judgment step which judges whether an anaphor exists in an input statement, and the anaphor which exists in an input statement based on the auxiliary information about the verb contained in the input statement, It is characterized by recording the program equipped with the antecedent decision step which determines the antecedent to which an anaphor points, and the analysis step which performs syntax analysis or the semantic analysis of an input statement using the antecedent determined in the antecedent decision step based on the attribute of an anaphor.

[0019] In a program, the basic sentence which is the unit made applicable [of a case frame] to generation is generated by the 1st natural-language-processing equipment of this invention and the natural-language-processing approach, and the list from the morphological analysis result of corpus data, and an unnecessary vocabulary is deleted from the basic sentence by generation of a case frame. Furthermore, about the verb in the basic sentence from which the unnecessary vocabulary was deleted, a case frame is generated, the subcategorization information and argument structure information on the verb are generated based on the case frame about the same verb, and it is outputted as auxiliary information.

[0020] It sets to a program at the 2nd natural-language-processing equipment of this invention and the natural-language-processing approach, and a list. From an auxiliary information storage means by which

the auxiliary information which consists of the subcategorization information and argument structure information on verbal is memorized at least While the auxiliary information about the verb contained in an input statement is retrieved, the attribute of the anaphor to which it is judged whether an anaphor exists in an input statement, and it exists in an input statement is recognized based on the auxiliary information about the verb contained in the input statement. And the antecedent to which an anaphor points is determined based on the attribute of an anaphor, and syntax analysis or the semantic analysis of an input statement is performed using the antecedent.

[0021]

[Embodiment of the Invention] Drawing 1 shows the example of a configuration of the gestalt of 1 operation of the natural-language-processing equipment which applied this invention.

[0022] This natural-language-processing equipment constitutes the auxiliary information generation equipment which searches for the auxiliary information which assists syntax analysis and the semantic analysis of natural language from a lot of corpus data.

[0023] That is, from a lot of corpus data, the natural-language-processing equipment as auxiliary information generation equipment of drawing 1 generates the case frame about a verb, and generates further the auxiliary information which includes the subcategorization information (subcategorization) and argument structure information (argument structure) on verbal from the case frame.

[0024] Here For example, Hiraoka crown 2 and Yuji Matsumoto (1994) "clustering of case frame acquisition [of the verb from a corpus], and noun" Information Processing Society of Japan, A natural-language-processing seminar, NL-104, and Masahiko Haruno (1995) "verb case frame study using principle of minimum generalization and Occam" Information Processing Society of Japan, A natural-language-processing seminar, NL-108, **** and Naoki Abe (1996) "Learning Dependencies between Case Frame Slots" Information Processing Society of Japan, Although the automatic generation method of the case frame for drawing up the dictionary called a thesaurus including homonymy relation information to a natural-language-processing seminar and NL-116 is indicated The case frame generated in the auxiliary information generation equipment of drawing 1 is a point aiming at creation of auxiliary information including subcategorization information and argument structure information, and differs from generating a case frame in order to create a thesaurus.

[0025] Moreover, the subcategorization information which constitutes auxiliary information For example HPSG () [Head-DrivenPhrase] Structure Grammar-C.Pollard & I.Sag (1996) Head-Driven Phrase Structure Grammar.CSLI & University of Chicago Press, JPSG () [Japanese Phrase] Structure Grammar- T. Gunji & KHasida (1998) Topics in Constraint-Based Grammar ofJapanese. Kluwer Academic Publishers; It is what bears an important role in the general-purpose natural-language-processing theory indicated by the Takao Gunji "research of the escape which took in concept of continuous volume of syntax based on constraint" (Heisei 12) Ministry of Education research result report etc. They are the following information.

[0026] That is, although a verb requires the constituent which has a certain specific structure and specific syntactic and semantic function, it is called subcategorization (subcategorization) to classify a verb on the basis of the constituent which the verb requires. like "being a restaurant and having eaten Japanese noodles with chopsticks", a verb "it eats" needs a noun phrase (Japanese noodles + "***") as a constituent, and, specifically, is further accompanied by the noun phrase (restaurant + -- "-- it is -- ") showing a location, and the noun phrase (chopsticks + -- "-- it is -- ") showing a means as a constituent if needed. Thus, subcategorization classifies a verb on the basis of the constituent which a verb needs, and the information about the constituent used as the criteria which classify a verb according to subcategorization is subcategorization information.

[0027] Furthermore, the constituent by which a verb is inevitably accompanied or it is accompanied if needed appears in what kind of location, and the argument structure information which constitutes auxiliary information means the information that what kind of semantic role is borne etc.

[0028] The auxiliary information generation equipment of drawing 1 consists of the corpus database 1, the pretreatment section 2, a case frame database 3, the case frame processing section 4, and an auxiliary information database 5.

[0029] The corpus database 1 has memorized a lot of corpus data. In addition, as corpus data, sentences, such as a newspaper article, are employable, for example.

[0030] The pretreatment section 2 consists of the morphological analysis section 11, the basic sentence pattern extract section 12, a cutout 13, and the case frame generation section 14, and performs processing which generates a case frame from a lot of corpus data memorized by the corpus database 1 as pretreatment which generates auxiliary information.

[0031] That is, the morphological analysis section 11 reads corpus data from the corpus database 1, and performs morphological analysis. And the morphological analysis section 11 supplies the morphological analysis result of corpus data to the basic sentence pattern extract section 12 and the case frame generation section 14. In addition, the morphological analysis result by the morphological analysis section 11 can be referred to now in the case frame processing section 4 mentioned later if needed.

[0032] From the morphological analysis result of the corpus data supplied from the morphological analysis section 11, the basic sentence pattern extract section 12 generates the basic sentence which is the unit made applicable [of a case frame] to generation (extract), and supplies it to a cutout 13. Namely, in principle, among the morphological analysis results which the morphological analysis section 11 outputs, the basic sentence pattern extract section 12 extracts from the next morpheme of a period to the morpheme in front of a period as a basic sentence, and supplies it to a cutout 13.

[0033] A cutout 13 deletes an unnecessary vocabulary from the basic sentence supplied from the basic sentence pattern extract section 12 to generation of a case frame, and supplies it to the case frame generation section 14.

[0034] Referring to the morphological analysis result of the corpus data supplied from the morphological analysis section 11 if needed, about the verb in the basic sentence supplied from a cutout 13, the case frame generation section 14 generates a case frame, and supplies it to the case frame database 3.

[0035] The case frame database 3 memorizes the case frame supplied from the pretreatment section 2 (case frame generation section 14 to constitute).

[0036] The case frame processing section 4 consists of the case frame integrated section 21, the verb classification section 22, the subcategorization information generation section 23, the argument structure information generation section 24, and the auxiliary information generation section 25. While reading the case frame about the same verb from the case frame database 3 and classifying the verb based on the case frame about the same verb etc., the subcategorization information and argument structure information are generated, and it outputs as auxiliary information.

[0037] That is, the case frame integrated section 21 reads the case frame about the same verb from the case frame database 3, and is taken as the integrated case frame which unifies and mentions those case frames later. And the case frame integrated section 21 supplies integrated each frame about each verb to the verb classification section 22, the subcategorization information generation section 23, and the argument structure information generation section 24.

[0038] The verb classification section 22 classifies into either of the four classification, an intransitive verb, a **** verb, a transitive verb, or a double object transitive verb, the verb corresponding to the integrated case frame supplied from the case frame integrated section 21, and supplies the classification information showing the classification result to the subcategorization information generation section 23 and the auxiliary information generation section 25.

[0039] Based on the integrated case frame supplied from the case frame integrated section 21, and the classification information supplied from the verb classification section 22, the subcategorization information generation section 23 generates the subcategorization information on the verb corresponding to the integrated case frame, and supplies it to the argument structure information generation section 24 and the auxiliary information generation section 25.

[0040] Based on the integrated case frame supplied from the case frame integrated section 21, and the subcategorization information supplied from the subcategorization information generation section 23, the argument structure information generation section 24 generates the argument structure information on the verb corresponding to the integrated case frame, and supplies it to the auxiliary information

generation section 25.

[0041] About each verb, the auxiliary information generation section 25 matches the classification information supplied from the verb classification section 22, the subcategorization information supplied from the subcategorization information generation section 23, and the argument structure information supplied from the argument structure information generation section 24, considers as auxiliary information, and is supplied to the auxiliary information database 5.

[0042] The auxiliary information database 5 memorizes the auxiliary information about each verb supplied from the auxiliary information generation section 25.

[0043] Next, drawing 2 shows the example of the morphological analysis result outputted when the morphological analysis section 11 carries out morphological analysis of the corpus data.

[0044] In addition, drawing 2 shows the morphological analysis result about for example, corpus data "fruits within the prefecture were especially conspicuous in quantity, and the increase of 34% and elongation were conspicuous in an increase and the amount of money 18%."

[0045] A morphological analysis result consists of a header of a morpheme, reading (phoneme), and thesaurus information, and thesaurus information includes the functor-attribute (feature) (functor attribute) of a morpheme, and a semantic attribute (semantic attribute). Furthermore, thesaurus information also includes the original form of the verb, when a morpheme is a verb.

[0046] here -- drawing 2 -- setting -- the 1st morpheme -- "-- especially -- " -- CAT of the attribute [CAT Adverv] in thesaurus information expresses that it is an attribute tag showing a part of speech, therefore the information which continues after that is a part of speech. Adverv which continues after CAT expresses that a part of speech is an adverb.

[0047] moreover, a morpheme -- "-- especially -- " -- attribute [VAL in thesaurus information -- the information which especially VAL of] is an attribute tag showing the value (header) of a morpheme, therefore continues after that -- "-- especially -- " -- it expresses that it is a corresponding morpheme.

[0048] The attribute [CAT Noun] in the thesaurus information on the 2nd morpheme "fruits within the prefecture" expresses that a part of speech is a noun. Moreover, cl of the attribute [cl Compound=CN+CN] in the thesaurus information on a morpheme "fruits within the prefecture" expresses that it is an attribute tag showing a class, therefore the information which continues after that is a class. Compound=CN+CN which continues after cl expresses that a class is the compound noun which the general noun (CN) and the general noun (CN) combined. Furthermore, Sem of the attribute [Sem food] in the thesaurus information on a morpheme "fruits within the prefecture" expresses that it is an attribute tag showing semantics, therefore the information which continues after that is semantics. food which continues after Sem expresses that it is that as which a morpheme means food. The attribute [VAL within-the-prefecture fruits] in the thesaurus information on a morpheme "fruits within the prefecture" expresses that the thesaurus information is a thing corresponding to a morpheme "fruits within the prefecture."

[0049] the 3rd morpheme -- "-- attribute [CAT Case] in the thesaurus information on " Expressing that a part of speech is a particle (Case), an attribute [cl abstract] expresses that a class is a case-marking particle (abstract). Furthermore, it expresses that fx of an attribute [fx nominative] is an attribute tag showing the function (grammatical role) of a morpheme, therefore the function of an attribute [fx nominative] is a nominative case (nominative). an attribute [VAL] -- the thesaurus information -- a morpheme -- "-- it expresses that it is a thing corresponding to ".

[0050] The attribute [CAT Noun] in the thesaurus information on the 4th morpheme "quantity" expresses that a part of speech is a noun, and an attribute [cl CNoun] expresses that a class is a general noun (CNoun). An attribute [Sem amount] expresses that it is that as which a morpheme "quantity" means an amount (amount), and an attribute [VAL quantity] expresses that the thesaurus information is a thing corresponding to a morpheme "quantity."

[0051] the 5th morpheme -- "-- it is -- " -- the attribute [CAT Case] in thesaurus information expresses that a part of speech is a particle, and an attribute [cl lexical] expresses that a class is a non-case-marking particle (lexical). the function of an attribute [fx instrument] is an instrument (instrument) -- expressing - an attribute [VAL] -- thesaurus information -- a morpheme -- "-- it is -- " -- it expresses that it is a

corresponding thing.

[0052] The attribute [CAT Noun] in the thesaurus information on the 6th morpheme "the increase of 18%" expresses that a part of speech is a noun, and an attribute [cl Compound=Num+Classifier+suf] expresses that a class is the compound (noun) which consists of a numeral (Num), a numerative (Classifier), and a suffix (suf). An attribute [Sem increase] expresses that it is that as which a morpheme "the increase of 18%" means an increment (increase), and an attribute [the increase of VAL18%] expresses that thesaurus information is a thing corresponding to a morpheme "the increase of 18%."

[0053] the 7th morpheme -- "-- the attribute [CAT Punctuation] in the thesaurus information on " -- a morpheme -- "-- expressing that " (part of speech) is a notation (Punctuation), an attribute [cl comma] expresses that a class is a comma (comma) (punctuation marks). an attribute [VAL] -- thesaurus information -- a morpheme -- "-- it expresses that it is a thing corresponding to ".

[0054] The attribute [CAT Noun] in the thesaurus information on the 8th morpheme "the amount of money" expresses that a part of speech is a noun, and an attribute [cl CNoun] expresses that a class is a general noun. An attribute [Sem money] expresses that it is that as which a morpheme "the amount of money" means money (money), and an attribute [the VAL amount of money] expresses that thesaurus information is a thing corresponding to a morpheme "the amount of money."

[0055] the 9th morpheme -- "-- it is -- " -- thesaurus information -- the 5th morpheme -- "-- it is -- " -- it is the same as that of a thing.

[0056] The thesaurus information on the 10th morpheme "the increase of 34%" is attribute [VAL. Except for 34% increase], it is the same as that of the thesaurus information on the 6th morpheme "the increase of 18%."

[0057] the 11th morpheme -- "-- ** -- " -- the attribute [CAT Complementizer] in thesaurus information expresses that a part of speech is a particle (Complementizer) which takes a complement sentence, and an attribute [cl proposition] expresses that a class is the citation (proposition) of a sentence. an attribute [VAL] -- thesaurus information -- a morpheme -- "-- ** -- " -- it expresses that it is a corresponding thing.

[0058] The attribute [CAT Noun] in the thesaurus information on the 12th morpheme "elongation" expresses that a part of speech is a noun, and an attribute [cl CNoun] expresses that a class is a general noun. An attribute [Sem increase] means that a morpheme "elongation" means an increment, and an attribute [VAL elongation] expresses that thesaurus information is a thing corresponding to a morpheme "elongation."

[0059] the 13th morpheme -- "-- the thesaurus information on " -- the 3rd morpheme -- "-- it is the same as that of the thing of ".

[0060] The attribute [CAT Verb] in the thesaurus information on the 14th morpheme "it was conspicuous" expresses that a part of speech is a verb (Verb), and an attribute [cl active] expresses that a class is activity (active). fm of an attribute [fm finite] is an attribute tag showing form, and an attribute [fm finite] expresses that form is a form (finite) accompanied by tense. Conj of an attribute [Conj (cl 2) (fm aff-past) (Polarity aff) (Ts past)] is an attribute tag showing an activity, and an attribute (cl 2) expresses that an activity is an activity of a class 2 (cl 2) (Stem it is conspicuous). Here, in the morphological analysis section 11, the class division of the verbal activity is carried out at some classes, and the activity of a class 2 means that the verbal original form finishes with a consonant. An attribute (Stem it is conspicuous) expresses that the original form (Stem) of a morpheme "it was conspicuous" is "conspicuous." In addition, Stem is an attribute tag showing the verbal original form. An attribute (fm aff-past) expresses that the form (fm) of a morpheme "it was conspicuous" is affirmation (aff (affirmation)), and is the past (past), and an attribute (Polarity aff) expresses that the polarity (Polarity) of a morpheme "it was conspicuous" is affirmation (aff). An attribute (Tspast) expresses that the tense (Ts) of a morpheme "it was conspicuous" is the past (past). Style of an attribute [Style (cl plain) (fm zero)] is an attribute tag showing a style (style), and an attribute (cl plain) expresses that the class (cl) of a style is an un-polite form (being the so-called -- "-- it is -- it is not measure tone"). The form (fm) of a style expresses that only the original form is (zero), and an attribute (fm zero) is attribute [VAL. Conspicuous] expresses that thesaurus information is a thing corresponding to a morpheme "it was

conspicuous."

[0061] the 15th morpheme -- ". -- " -- the attribute [CAT Punctuation] in thesaurus information -- a morpheme -- ". -- " (part of speech) -- expressing that it is a notation (Punctuation), an attribute [cl period] expresses that a class is a period (period) (period). Attribute [VAL.] ** and thesaurus information -- a morpheme -- ". -- " -- it expresses that it is a corresponding thing.

[0062] Next, drawing 3 shows the example of the vocabulary which a cutout 13 deletes from the basic sentence supplied from the basic sentence pattern extract section 12 as an unnecessary vocabulary (suitably henceforth an unnecessary vocabulary) to generation of a case frame.

[0063] A cutout 13 deletes eight kinds of following vocabularies from a basic sentence as an unnecessary vocabulary.

[0064] That is, a cutout 13 deletes an adverb from a basic sentence as an unnecessary vocabulary to the 1st. An adverb is detectable by searching the morpheme from which thesaurus information is {[CAT Adverb]} from a morphological analysis result, as shown in drawing 3 (A).

[0065] A cutout 13 deletes the noun + particle "***" "summer" etc. from a basic sentence as an unnecessary vocabulary to the 2nd, for example. For a noun + particle "***", as shown in drawing 3 (B), a morphological analysis result to thesaurus information is {[CAT Noun]... It is detectable by searching the part which the morpheme used as} and the morpheme used as {[CAT Case], [cl abstract], [fx genitive], and [VAL]} are following.

[0066] in addition, drawing 3 -- setting (also setting to drawing 5 mentioned later the same) -- parenthesis {} -- it means that, as for inner ..., other attributes may be described.

[0067] A cutout 13 deletes the noun + particle + particle "***" "Japan" etc. from a basic sentence as an unnecessary vocabulary to the 3rd, for example. For a noun + particle + particle "***", as shown in drawing 3 (C), a morphological analysis result to thesaurus information is {[CAT Noun]... It is the morpheme and {[CAT Case] used as}... It is detectable by searching the part which the morpheme used as} and the morpheme used as {[CAT Case], [cl abstract], [fx genitive], and [VAL]} are following.

[0068] A cutout 13 deletes an adjective from a basic sentence as an unnecessary vocabulary to the 4th. For an adjective, as shown in drawing 3 (D), a morphological analysis result to thesaurus information is {[CAT Adjective] and [cl stative]... It is detectable by searching the morpheme used as}. In addition, an attribute [CAT Adjective] expresses that a part of speech is an adjective (Adjective), and an attribute [cl stative] expresses that a class is a condition (stative).

[0069] noun (adjective verb word stem) + [cutout / 13 / 5th / the / sentence / basic] for example "it is decisive" etc. -- "-- " is deleted as an unnecessary vocabulary. noun (adjective verb word stem) + -- "-- " is shown in drawing 3 (E) -- as -- a morphological analysis result to thesaurus information -- {[CAT Noun] ... the morpheme used as}, and {[CAT Verb] and [cl copula] ... it is detectable by searching the part which the morpheme used as} [[VAL]] is following. In addition, an attribute [cl copula] expresses that a class is a copula.

[0070] A cutout 13 deletes a noun + say postposition from a basic sentence as an unnecessary vocabulary "to works" etc. to the 6th, for example. For a noun + postposition, as shown in drawing 3 (F), a morphological analysis result to thesaurus information is {[CAT Noun]... It is [the morpheme used as}, and] {[CAT Postposition]... It is detectable by searching the part which the morpheme used as} is following. In addition, an attribute [CAT Postposition] expresses that a part of speech is a postposition (Postposition).

[0071] A cutout 13 deletes the part enclosed in the parenthesis from a basic sentence as an unnecessary vocabulary to the 7th. a parenthesis -- surrounding -- having had -- a part -- drawing 3 -- (-- G --) -- being shown -- as -- morphological analysis -- a result -- from -- a thesaurus -- information -- {-- [-- CAT Punctuation --] -- [-- cl L -] --} -- becoming -- **** -- a morpheme -- from -- {-- [-- CAT Punctuation --] -- [-- cl R -] --} -- becoming -- **** -- a morpheme -- up to -- a part -- searching -- things -- being detectable . In addition, an attribute [cl L-] expresses that a class is a parenthesis (it, for example, expresses that they are "(etc.), and a class is closing parenthesis (for example,")" etc. for an attribute [cl R-]).

[0072] A cutout 13 deletes the partial + particle "***" enclosed in the parenthesis from a basic sentence

as an unnecessary vocabulary to the 8th. As shown in drawing 3 (H), the partial + particle "***" enclosed in the parenthesis from a morphological analysis result The part to the morpheme to which thesaurus information is {[CAT Punctuation] and [cl R-]} from the morpheme used as [CAT Punctuation] and [cl L-]}, It is detectable by searching after that the morpheme from which thesaurus information serves as {[CAT Case], [cl abstract], [fx genitive], and [VAL]}.

[0073] In a cutout 13, eight kinds of above vocabularies are deleted from a basic sentence as an unnecessary vocabulary.

[0074] About the corpus data "fruits within the prefecture were especially conspicuous in quantity, and the increase of 34% and elongation were conspicuous in an increase and the amount of money 18%." which followed, for example, were mentioned above, the following basic sentences are outputted from a cutout 13.

[0075] That is, in the basic sentence pattern extract section 12, "fruits within the prefecture were especially conspicuous in quantity, and the increase of 34% and elongation were conspicuous about corpus data "fruits within the prefecture were especially conspicuous in quantity, and the increase of 34% and elongation were conspicuous in an increase and the amount of money 18%.", in an increase and the amount of money 18%" is extracted as a basic sentence. [which removed the period from the corpus data] and the morpheme which corresponds to the part of speech of drawing 3 (A) being an adverb in a cutout 13 from "fruits within the prefecture were especially conspicuous in quantity, and the increase of 34% and elongation were conspicuous in an increase and the amount of money 18%" -- "-- especially -- " -- it is deleted and "fruits within the prefecture were especially conspicuous in quantity, and the increase of 34% and elongation were conspicuous in an increase and the amount of money 18%" is outputted.

[0076] therefore, the morpheme which is an adverb as the morphological analysis result of the corpus data "fruits within the prefecture were especially conspicuous in quantity, and the increase of 34% and elongation were conspicuous in an increase and the amount of money 18%." shown in drawing 2 is shown in drawing 4 in a cutout 13 -- "-- especially -- " -- the morpheme which is a period -- ". -- " -- it becomes a thing without the related information and is outputted.

[0077] Next, although the case frame generation section 14 generates a case frame about the verb in the basic sentence which a cutout 13 outputs, generation of this case frame is performed, using the "criteria form" of the verb contained in a basic sentence as a header of a case frame. That is, a case frame consists of a header of the verb with which the case frame expresses whether it is a thing about what kind of verb, and information about the particle by which the verb is accompanied in a basic sentence, and a verbal criteria form is used as a header of a case frame.

[0078] Here, as it is indicated in drawing 5 as the verbal criteria form used as the header of a case frame, it defines.

[0079] That is, in principle except for three exceptions explained below, the original form of the verb contained in a basic sentence turns into a criteria form of the verb. As shown in drawing 5 (A), the morpheme "it is conspicuous" which is a verb, and when "it was conspicuous" is included by the basic sentence, specifically, the original form "it is conspicuous" turns into a criteria form at it.

[0080] In addition, since the verbal original form is described with the Stem attribute tag in the thesaurus information on a morphological analysis result as drawing 2 explained, it can be recognized by referring to thesaurus information.

[0081] Next, as the 1st exception, when the SA strange noun + verb "it carries out" is contained in the basic sentence, not the verbal "it carries out" original form but a SA strange noun + verb "it carries out" serves as a verbal criteria form.

[0082] therefore -- for example, it is shown in drawing 5 (B) -- as -- the thesaurus information on a morphological analysis result -- {[CAT Noun] and [cl Vnoun] ... the morpheme "application" used as [VAL application]}, {[CAT Verb] and [cl active], and [fm finite] ... (Stem it carries out) (fm aff-non-past) ... when the morpheme "it carries out" used as} [carried out VAL] continues, it considers as the criteria form of the verb "to apply". in addition, an attribute [cl Vnoun] expresses that a class is a SA strange noun (Vnoun), and an attribute (fm aff-non-past) is affirmation (aff), and the form (fm) of a

morpheme "it carries out" is not the past (non-past) -- it is -- things are expressed.

[0083] As the 2nd exception, the original form of the verb of the beginning of the two verbs which the verb of the beginning of them follows when two verbs have two attributes, [fm infinite] and (pres.participle), in thesaurus information continuously in a basic sentence turns into a verbal criteria form. In addition, form expresses that it is a form (infinite) without tense, and an attribute [fm infinite] expresses that an attribute (pres.participle) is the present participle (presentparticiple).

[0084] therefore -- for example, the morpheme which has two attributes, [fm infinite] and (pres.participle), in a basic sentence as shown in drawing 5 (C) -- "-- expecting -- " -- then -- the case where "it expects" exists when there is a morpheme "it is" -- a morpheme -- "-- expecting -- " -- the original form "it expects" is made into a verbal criteria form.

[0085] When the verb whose original form is "carrying out" is contained in a basic sentence as the 3rd exception and a SA strange noun is just before the verb, SA strange noun + "it carries out" becomes a verbal criteria form.

[0086] As it follows, for example, is shown in drawing 5 (D), the thesaurus information on a morphological analysis result {[CAT Noun], [cl Vnoun] ... The morpheme used as [VAL expansion]} "expansion", {-- -- CAT Verb} ... [fm infinite] ... (Stem it carries out) (fm pres.participle) ... [-- the morpheme which carries out VAL and has become]} -- "-- carrying out -- " -- And {[CAT Verb] ... [fm finite] ...(Stem it is)... When the morpheme "it is" used as} [which is VAL] is continuing, it is carried out to SA strange noun "expansion" + "it carries out", i.e., the criteria form of the verb "to develop."

[0087] Next, drawing 6 shows the case frame which the case frame generation section 14 creates.

[0088] four case frames by which drawing 6 was generated from four basic sentences about the verb "it is conspicuous", respectively -- {-- conspicuous C_FRAME: -- [instrument] , -- [increase]} -- {-- conspicuous C_FRAME: -- [thing]} -- {-- conspicuous C_FRAME: and [proposition] , -- [thing]} -- {-- [locative] , shows [increase]} to [instrument] by conspicuous C_FRAME:.

[0089] The character string of the head of a case frame expresses the header of the verb corresponding to that case frame, and the criteria form of the verb explained by drawing 5 is used as a header of this verb.

[0090] Moreover, C_FRAME in a case frame is a tag showing a particle (case-marking particle), and the particle which the verb which is the header has taken in the basic sentence is described by after that. In addition, one or more particles can be described to a case frame.

[0091] furthermore -- although parenthesis [] is described just behind the particle in a case frame -- this parenthesis [] -- inside, the semantics of the function of that particle or the morpheme in front of that particle is described as an attribute of that particle. In addition, the function of a particle can be recognized by searching fx attribute tag of the thesaurus information in a morphological analysis result, and the semantics of the morpheme in front of a particle can be recognized by searching the Sem attribute tag of thesaurus information.

[0092] the case frame [in / here / drawing 6] of the 1st line -- {-- it is generated by performing case frame generation processing of drawing 12 which the case frame generation section 14 mentions [[instrument] , / [increase]}] later by conspicuous C_FRAME: about above-mentioned corpus data "fruits within the prefecture were especially conspicuous in quantity, and the increase of 34% and elongation were conspicuous in an increase and the amount of money 18%."

[0093] Next, drawing 7 shows the integrated case frame generated when the case frame integrated section 21 unifies the case frame about the same verb.

[0094] For example, when four case frames as shown in drawing 6 are obtained about the verb (criteria form) "it is conspicuous", an integrated case frame as shown in drawing 7 is generated about a verb "it is conspicuous" by unifying the four case frames.

[0095] That is, in this case, the case frame integrated section 21 arranges the header "it is conspicuous" of the verb to four case frames about a verb "it is conspicuous" as a header of an integrated case frame, and arranges reading of that verb continuously. In addition, verbal reading is recognized because the case frame integrated section 21 refers to the morphological analysis result of the morphological analysis section 11.

[0096] Furthermore, the case frame integrated section 21 asks for the particle of four case frames, and

[0097] namely, -- four case frames shown in drawing 6 -- "-- it is -- " -- "-- " -- "-- ** -- " -- "-- alike -- " -- since four kinds of particles exist -- the case frame integrated section 21 -- four kinds of this particle -- "-- it is -- " -- "-- " -- "-- ** -- " -- "-- alike -- " -- it arranges after Tag subcat. furthermore, four case frames of drawing 6 -- setting -- a particle -- "-- it is -- " -- ***** -- the particle in an integrated case frame since only an attribute [instrument] exists -- "-- it is -- " -- behind, the attribute [instrument] is arranged. moreover, four case frames of drawing 6 -- setting -- a particle (case-marking particle) -- "-- the particle in an integrated case frame since two kinds, an attribute [increase] and [thing], exist about " - - "-- after ", the two attributes [increase] and [thing(s)] are arranged. furthermore, four case frames of drawing 6 -- setting -- a particle -- "-- ** -- " -- ***** -- the particle in an integrated case frame since only an attribute [proposition] exists -- "-- ** -- " -- behind, the attribute [proposition] is arranged. moreover, four case frames of drawing 6 -- setting -- a particle -- "-- alike -- " -- ***** -- the particle in an integrated case frame since only an attribute [locative] exists -- "-- alike -- " -- behind, the attribute [locative] is arranged.

[0099] Drawing 8 shows the auxiliary information about a verb "it is conspicuous", and a verbal header (verbal criteria form) and verbal "it is conspicuous" reading are arranged like the integrated case frame indicated to be a head to the 2nd at drawing 7 .

[0101] In auxiliary information, the subcategorization information on verbal is arranged after classification information. Subcategorization information is described in the form of <SUBCAT:NP [nom]> with the tag SUBCAT showing being subcategorization information, as shown in drawing 8 . In addition, NP expresses a noun phrase and [nom] expresses a nominative case. And it means that subcategorization information <SUBCAT:NP [nom]> is inevitably accompanied by the noun phrase used as a nominative case. This subcategorization information is supplied to the auxiliary information generation section 25 from the subcategorization information generation section 23.

[0103] brace {} after the main information -- inner description expresses the attribute (thesaurus) of the main information. thing and increase of {thing/increase} in drawing 8 express an increment as an object, respectively, therefore an attribute {thing/increase} expresses that it is what the object which the main information "Theme" expresses expresses an object or an increment to.

<http://www4.ipdl.inpit.go.jp/cgi-bin/transweb.cgi>

by the expression with which a verb "it is conspicuous" expresses an instrument, the expression showing a location, and the expression indicating a sentence if needed.

[0105] Next, with reference to the flow chart of drawing 9, the auxiliary information generation processing which generates the auxiliary information as natural language processing which the auxiliary information generation equipment of drawing 1 performs is explained.

[0106] First, in step S1, the morphological analysis section 11 reads a lot of corpus data memorized by the corpus database 1 one by one, and performs morphological analysis about each corpus data. The morphological analysis result obtained when the morphological analysis section 11 performs morphological analysis about each corpus data is supplied to the basic sentence pattern extract section 12 and the case frame generation section 14, and a list at the case frame processing section 4.

[0107] Then, it progresses to step S2, and from the morphological analysis result which each corpus data supplied from the morphological analysis section 11 attaches, the basic sentence pattern extract section 12 performs basic sentence pattern extract processing in which a basic sentence is extracted, it supplies the basic sentence obtained as a result to a cutout 13, and progresses to step S3. At step S3, unnecessary lexical deletion to which a cutout 13 deletes an unnecessary vocabulary from each basic sentence supplied from the basic sentence pattern extract section 12 is performed, the basic sentence which deleted the unnecessary vocabulary is supplied to the case frame generation section 14, and it progresses to step S4. In step S4, the case frame generation section 14 performs case frame generation processing which generates a case frame about the verb contained in the basic sentence about each basic sentence supplied from a cutout 13. Furthermore, the case frame generation section 14 supplies and stores in the case frame database 3 the case frame generated by the case frame generation processing, and progresses to step S5.

[0108] At step S5, from the case frame memorized by the case frame database 3, as the things about the same verb are collected and drawing 6 and drawing 7 explained, the case frame integrated section 21 unifies one or more case frames about the same verb, and generates an integrated case frame. And the case frame integrated section 21 supplies an integrated case frame to the verb classification section 22, the subcategorization information generation section 23, and the argument structure information generation section 24, and progresses to step S6.

[0109] At step S6, based on the integrated case frame supplied from the case frame integrated section 21, the verb classification section 22 classifies the verb corresponding to each integrated case frame into an intransitive verb, a **** verb, a transitive verb, or a double object transitive verb, and performs verb classification processing which outputs the classification information showing the classification result. Furthermore, at step S6, subcategorization information generation processing in which the subcategorization information generation section 23 generates and outputs the subcategorization information on the verb corresponding to each integrated case frame based on the integrated case frame supplied from the case frame integrated section 21 and the classification information supplied from the verb classification section 22 is performed. Moreover, at step S6, argument structure information generation processing in which the argument structure information generation section 24 generates and outputs the argument structure information on the verb corresponding to each integrated case frame based on the integrated case frame supplied from the case frame integrated section 21 and the subcategorization information supplied from the subcategorization information generation section 23 is performed.

[0110] Then, it progresses to step S7 and the auxiliary information generation section 25 generates auxiliary information as shown in drawing 8 about the verb corresponding to each integrated case frame using the classification information supplied from the verb classification section 22, the subcategorization information supplied from the subcategorization information generation section 23, and the argument structure information supplied from the argument structure information generation section 24. Furthermore, the auxiliary information generation section 25 supplies and stores auxiliary information in the auxiliary information database 5, and ends auxiliary information generation processing.

[0111] Next, with reference to the flow chart of drawing 10, the basic sentence pattern extract

processing which the basic sentence pattern extract section 12 of drawing 1 performs at step S2 of drawing 9 is explained.

[0112] In step S11, the basic sentence pattern extract section 12 uses as attention corpus data the oldest thing that has not been made into the object of processing yet among the corpus data with which the morphological analysis result was obtained in the morphological analysis section 11 while clearing the buffer (not shown) which it has. And it progresses to step S12, and considering what that has not read the morpheme of attention corpus data yet and is more close to a beginning of a sentence as an attention morpheme, the basic sentence pattern extract section 12 reads the morphological analysis result, and progresses to step S13. At step S13, the basic sentence pattern extract section 12 judges the morphological analysis result by referring to for whether an attention morpheme is a period.

[0113] the case where it is judged with an attention morpheme not being a period in step S13 -- step S14 -- progressing -- the basic sentence pattern extract section 12 -- the morphological analysis result of an attention morpheme -- among those, the buffer which it has is made to carry out additional storage, and the same processing is hereafter repeated for the following morpheme which is step S12 with return and a now attention morpheme as a new attention morpheme.

[0114] Moreover, in step S13, by progressing to step S15, when are judged with an attention morpheme being a period, and the basic sentence pattern extract section 12 refers to the buffer to build in, it judges whether the morpheme in front of the period which is an attention morpheme (or the first verb which exists before a period) is a verb accompanied by tense. In step S15, when it judges that the morpheme in front of the period which is an attention morpheme is not a verb accompanied by tense, steps S16 and S17 are skipped, and it progresses to step S18.

[0115] Moreover, in step S15, when it judges that the morpheme in front of the period which is an attention morpheme is a verb accompanied by tense, it progresses to step S16 and the basic sentence pattern extract section 12 judges whether the verb (morphological analysis result) accompanied by tense is memorized by the buffer to build in other than the morpheme in front of the period which is an attention morpheme.

[0116] In step S16, when judged with the verb accompanied by tense being memorized in addition to the morpheme in front of the period which is an attention morpheme by the buffer which the basic sentence pattern extract section 12 builds in, step S17 is skipped and it progresses to it at step S18.

[0117] On the other hand in step S16, to the buffer which the basic sentence pattern extract section 12 builds in When judged with the verb accompanied by tense not being memorized other than the morpheme in front of the period which is an attention morpheme, it progresses to step S17. The basic sentence pattern extract section 12 The sequence of the morpheme (analysis result) memorized by the buffer to build in is extracted as a basic sentence (read-out), is supplied to a cutout 13, and it progresses to step S18.

[0118] At step S18, the basic sentence pattern extract section 12 judges whether there are still any corpus data which have not been used as attention corpus data. In step S18, when judged with there being still corpus data which have not been used as attention corpus data, return and one of the corpus data which have not been used as attention corpus data yet are newly used as attention corpus data, and the same processing is hereafter repeated by step S11.

[0119] Moreover, in step S18, when judged with there being still no corpus data which have not been used as attention corpus data, basic sentence pattern extract processing is ended.

[0120] According to the above basic sentence pattern extract processings, it is a morphological train from the morpheme just behind a period to the morpheme in front of the following period, and that (fundamentally simple sentence) in which only one contains the verb accompanied by tense is extracted as a basic sentence.

[0121] Next, with reference to the flow chart of drawing 11, the unnecessary lexical deletion which the cutout 13 of drawing 1 performs at step S3 of drawing 9 is explained.

[0122] A cutout 13 first still sets to Variable N the number of the morphemes which constitute the attention basic sentence as an attention basic sentence for one of those which are considering as the attention basic sentence and are not among the basic sentences supplied from the basic sentence pattern

extract section 12 in step S21.

[0123] And a cutout 13 progresses to step S22, all initializes the variables i and j which count the morpheme of a basic sentence to 1, and progresses to step S23.

[0124] At step S23, a cutout 13 sets the morphological train from the i-th morpheme to [from the head of an attention basic sentence] the j-th morpheme to Variable String, and progresses to step S24.

[0125] At step S24, a cutout 13 judges whether the morphological train (or morpheme) set to Variable String corresponds to deletion conditions.

[0126] Here, the case where it corresponds to deletion conditions means corresponding to either of the unnecessary vocabularies explained by drawing 3 .

[0127] In step S24, when judged with the morphological train set to Variable String not corresponding to deletion conditions, step S25 is skipped and it progresses to step S26. Moreover, in step S24, when judged with the morphological train set to Variable String corresponding to deletion conditions, it progresses to step S25, and a cutout 13 buffers the morphological train set to the buffer (not shown) to build in by Variable String as an object for deletion, and progresses to step S26.

[0128] At step S26, a cutout 13 judges whether it is equal to several Ns of the morpheme from which Variable j constitutes an attention basic sentence. In step S26, when it judges that Variable j is not equal to N, it progresses to step S27, and only 1 increments Variable j and a cutout 13 repeats the same processing return and the following to step S23.

[0129] Moreover, in step S26, when it judges that Variable j is equal to N, it progresses to step S28 and judges whether a cutout 13 has Variable i equal to N. In step S28, when it judges that Variable i is not equal to N, it progresses to step S29, and a cutout 13 sets the value set to Variable j by Variable i, and repeats the same processing return and the following to step S23 while only 1 increments Variable i.

[0130] When it judges that Variable i is equal to N in step S28 on the other hand, About the morpheme and the morphological train of arbitration which constitute a basic sentence, when it is judged that it is an unnecessary vocabulary, it progresses to step S30. Namely, a cutout 13 The morpheme and the morphological train which are memorized as an object for deletion by the buffer which it has are deleted from an attention basic sentence, the case frame generation section 14 is supplied, and it progresses to step S31.

[0131] At step S31, a cutout 13 judges whether there is still any basic sentence which has not been made into the attention basic sentence. In step S31, when judged with there being still a basic sentence which has not been made into the attention basic sentence, to step S21, return and a cutout 13 still make one of the basic sentences which have not been made into the attention basic sentence a new attention basic sentence, and repeat the same processing hereafter.

[0132] Moreover, in step S31, when judged with there being still no basic sentence which has not been made into the attention basic sentence, unnecessary lexical deletion is ended.

[0133] Next, with reference to the flow chart of drawing 12 , the case frame generation processing which the case frame generation section 14 of drawing 1 performs at step S5 of drawing 9 is explained.

[0134] The case frame generation section 14 first still describes the criteria form of the verb (suitably henceforth an attention verb) contained in the attention basic sentence as an attention basic sentence in one of those which are considering as the attention basic sentence and are not as a header of the case frame about the attention verb among the basic sentences supplied from a cutout 13 in step S41.

[0135] And the case frame generation section 14 progresses to step S42, initializes to 1 the variable i which counts the morpheme of a basic sentence, and progresses to step S43.

[0136] At step S43, the case frame generation section 14 sets the i-th morpheme to Variable String from the last of an attention basic sentence, and progresses to step S44.

[0137] At step S44, the case frame generation section 14 judges the thesaurus information on the morphological analysis result (drawing 2) by referring to for whether the morpheme set to Variable String is a particle.

[0138] In step S44, when judged with the morpheme set to Variable String not being a particle, steps S45 and S46 are skipped, and it progresses to step S47.

[0139] Moreover, in step S44, when judged with the morpheme set to Variable String being a particle, it

progresses to step S45, and the case frame generation section 14 describes the attribute to be the particle set to Variable String to the case frame about an attention verb, and progresses to step S46. In addition, the case frame generation section 14 is recognized by referring to the thesaurus information on the morphological analysis result according the attribute of a particle to the morphological analysis section 11.

[0140] the particle by which the case frame generation section 14 is set to Variable String at step S46 -- from the last of an attention basic sentence -- counting -- the 1st "***" or the 2nd -- "-- " -- "-- alike -- " -- or it judges whether it corresponds to either of the "***."

[0141] the particle set to Variable String in step S46 -- from the last of an attention basic sentence -- counting -- the 1st "***" and the 2nd -- "-- " and the 2nd -- "-- alike -- " -- or when judged with corresponding to either of the 2nd "***", step S47 is skipped and it progresses to step S49.

[0142] In step S46, moreover, the particle set to Variable String from the last of an attention basic sentence -- counting -- the 1st "***" and the 2nd -- "-- " -- the 2nd -- "-- alike -- " -- and when judged with corresponding to neither of the 2nd "***", it progresses to step S47 and judges whether the morpheme by which the case frame generation section 14 is set to Variable String is a morpheme of the head of an attention basic sentence.

[0143] In step S47, when it judges that the morpheme set to Variable String is not a morpheme of the head of an attention basic sentence, it progresses to step S48, and only 1 increments Variable i and the case frame generation section 14 repeats the same processing return and the following to step S43.

[0144] Moreover, in step S47, when it judges that the morpheme set to Variable String is a morpheme of the head of an attention basic sentence, it progresses to step S49 and the case frame generation section 14 judges whether there is still any basic sentence which has not been made into the attention basic sentence. In step S49, when judged with there being still a basic sentence which has not been made into the attention basic sentence, to step S41, return and the case frame generation section 14 still make one of the basic sentences which have not been made into the attention basic sentence a new attention basic sentence, and repeat the same processing hereafter.

[0145] Moreover, in step S49, when judged with there being still no basic sentence which has not been made into the attention basic sentence, case frame generation processing is ended.

[0146] According to the above case frame generation processings, it follows in the direction of a beginning of a sentence from the sentence end of the basic sentence which a cutout 13 outputs. Or it is described by the case frame about the particle which will appear by the time it arrives at either of the 2nd "***", and the verb by which the attribute is included in the basic sentence. the 1st "***" and the 2nd -- "-- " and the 2nd -- "-- alike -- " -- thereby A case frame as shown in drawing 6 is generated.

[0147] Next, with reference to the flow chart of drawing 13, the verb classification processing which the verb classification section 22 of drawing 1 performs at step S6 of drawing 9 is explained.

[0148] In step S61, among the integrated case frames which the case frame integrated section 21 outputs, although the verb classification section 22 has not considered as an attention integrated case frame, it makes one an attention integrated case frame, and it still reads subcategory information from the attention integrated case frame.

[0149] Here, the information described after a subcat tag is meant in the integrated case frame indicated to be subcategory information to drawing 7.

[0150] It progresses to step S62. Then, the verb classification section 22 Although an attention integrated case frame does not contain a case-marking particle "***" in the subcategory information a case-marking particle -- "-- " -- containing -- and the case-marking particle -- "-- the noun + case-marking particle which consists of nouns as " -- "-- it judges whether the conditions (suitably henceforth intransitive verb conditions) which the intransitive verb that " can become the agent (agent) of the verb corresponding to an attention integrated case frame fulfills are fulfilled.

[0151] here -- a noun + case-marking particle -- "-- actuation of the verb corresponding to an attention integrated case frame in " -- it can be judged by referring to the Sem tag showing the semantics of the thesaurus information in the morphological analysis result of the corpus data containing the verb whether it can become main.

[0152] In step S62, when it judges that an attention integrated case frame fulfills intransitive verb conditions, it progresses to step S63, and the verb classification section 22 classifies the verb (verb used as the header of an attention integrated case frame) corresponding to an attention integrated case frame into an intransitive verb, and it supplies to the subcategorization information generation section 23 and the auxiliary information generation section 25, and it progresses the classification information showing that to step S71.

[0153] When it judges that an attention integrated case frame does not fulfill intransitive verb conditions in step S62, it progresses to step S64. Moreover, the verb classification section 22 Although an attention integrated case frame does not contain a case-marking particle "***" in the subcategory information a case-marking particle -- "-- " -- containing -- and the case-marking particle -- "-- the noun + case-marking particle which consists of nouns as " -- "-- it judges whether the conditions (suitably henceforth **** verb conditions) which the **** verb that " cannot become the agent (agent) of the verb corresponding to an attention integrated case frame fulfills are fulfilled.

[0154] In step S64, when it judges that an attention integrated case frame fulfills **** verb conditions, it progresses to step S65, and the verb classification section 22 classifies the verb corresponding to an attention integrated case frame into a **** verb, supplies the classification information showing that to the subcategorization information generation section 23 and the auxiliary information generation section 25, and progresses to step S71.

[0155] moreover, the particle which needs an attention integrated case frame to take an indirect object although it progresses to step S66 and the verb classification section 22 contains a case-marking particle "***" in the subcategory information when it judges that an attention integrated case frame does not fulfill **** verb conditions in step S64 -- "-- alike -- " -- it judges whether the conditions (suitably henceforth transitive verb conditions) which the transitive verb of not containing fulfills are fulfilled.

[0156] In step S66, when it judges that an attention integrated case frame fulfills transitive verb conditions, it progresses to step S67, and the verb classification section 22 classifies the verb corresponding to an attention integrated case frame into a transitive verb, supplies the classification information showing that to the subcategorization information generation section 23 and the auxiliary information generation section 25, and progresses to step S71.

[0157] moreover, a particle required [as for the verb classification section 22] to progress to step S68 and for an attention integrated case frame take an indirect object further to the subcategory information including a case-marking particle "***" when it judges that an attention integrated case frame does not fulfill transitive verb conditions in step S66 -- "-- alike -- " -- it judges whether the conditions (suitably henceforth double object transitive verb conditions) which the double object transitive verb of containing fulfills are fulfilled.

[0158] In step S68, when it judges that an attention integrated case frame fulfills double object transitive verb conditions, it progresses to step S69, and the verb classification section 22 classifies the verb corresponding to an attention integrated case frame into a double object transitive verb, supplies the classification information showing that to the subcategorization information generation section 23 and the auxiliary information generation section 25, and progresses to step S71.

[0159] Moreover, in step S68, when it judges that an attention integrated case frame does not fulfill double object transitive verb conditions, it progresses to step S70, for example, error processing of excepting an attention integrated case frame from the processing object in the case frame processing section 4 is performed, and it progresses to step S71.

[0160] At step S71, it judges whether there is any integrated case frame which the verb classification section 22 has not made an attention integrated case frame yet. In step S71, when judged with there being still an integrated case frame which has not been made into the attention integrated case frame, to step S61, return and the verb classification section 22 still make one of the integrated case frames which have not been made into the attention integrated case frame a new attention integrated case frame, and repeat the same processing hereafter.

[0161] Moreover, in step S71, when judged with there being still no integrated case frame which has not been made into the attention integrated case frame, verb classification processing is ended.

[0162] Next, with reference to the flow chart of drawing 14, the subcategorization information generation processing which the subcategorization information generation section 23 of drawing 1 performs at step S6 of drawing 9 is explained.

[0163] First, among the integrated case frames which the case frame integrated section 21 outputs in step S81, although the subcategorization information generation section 23 has not considered as an attention integrated case frame, it receives one as an attention integrated case frame, and it still receives further the classification information which the verb classification section 22 outputs about the attention integrated case frame.

[0164] And it progresses to step S82 and the subcategorization information generation section 23 generates the subcategorization information on the verb corresponding to an attention integrated case frame based on an attention integrated case frame and its classification information.

[0165] Namely, the verb corresponding to an attention integrated case frame in the subcategorization information generation section 23 The attention verb recognizes any of an intransitive verb, a **** verb, a transitive verb, or the double object transitive verbs they are from the classification information on (calling it an attention verb suitably hereafter). The recognition result, From an attention integrated case frame, an attention verb recognizes the constituent by which it is accompanied inevitably (constraint is applied to the constituent by which the attention verb is accompanied inevitably by any of the four above-mentioned verbs attention verbs are under the constraint). From an attention integrated case frame, an attention verb recognizes the constituent by which it is accompanied inevitably. And the subcategorization information generation section 23 outputs the information about the constituent by which the attention verb is accompanied inevitably as subcategorization information to the argument structure information generation section 24 and the auxiliary information generation section 25.

[0166] As it followed, for example, having considered the case where it was presupposed that the integrated case frame about the verb "it is conspicuous" now shown in drawing 7 was made into the attention integrated case frame the verb "it is conspicuous" was first mentioned above, it is a **** verb and is inevitably accompanied by the noun phrase used as a nominative case. moreover, the case-marking particle which expresses a nominative case in the integrated case frame about the verb "it is conspicuous" shown in drawing 7 -- "-- only " exists and other case-marking particles do not exist. Then, NP [nom] which means being inevitably accompanied by the noun phrase used as a nominative case in the subcategorization information generation section 23 is generated as subcategorization information on verbal "it is conspicuous." In addition, as drawing 8 explained, NP expresses a noun phrase and [nom] expresses a nominative case.

[0167] Then, it progresses to step S83 and judges whether there is any integrated case frame which the subcategorization information generation section 23 has not made an attention integrated case frame yet. In step S83, when judged with there being still an integrated case frame which has not been made into the attention integrated case frame, to step S81, return and the subcategorization information generation section 23 still make one of the integrated case frames which have not been made into the attention integrated case frame a new attention integrated case frame, and repeat the same processing hereafter.

[0168] Moreover, in step S83, when judged with there being still no integrated case frame which has not been made into the attention integrated case frame, subcategorization information generation processing is ended.

[0169] Next, with reference to the flow chart of drawing 15, the argument structure information generation processing which the argument structure information generation section 24 of drawing 1 performs at step S6 of drawing 9 is explained.

[0170] First, among the integrated case frames which the case frame integrated section 21 outputs in step S91, although the argument structure information generation section 24 has not considered as an attention integrated case frame, it receives one as an attention integrated case frame, and it still receives further the subcategorization information which the subcategorization information generation section 23 outputs about the attention integrated case frame.

[0171] And it progresses to step S92 and the argument structure information generation section 24 recognizes the attribute to be an attention integrated case frame and the case-marking particle by which

the verb corresponding to an attention integrated case frame is inevitably accompanied based on the subcategorization information (indispensable).

[0172] That is, the argument structure information generation section 24 recognizes an indispensable case-marking particle for the attention verb from the subcategorization information on the verb (suitably henceforth an attention verb) corresponding to an attention integrated case frame, and recognizes the attribute of the case-marking particle from an attention integrated case frame further.

[0173] When the integrated case frame about the verb "it is conspicuous" which followed, for example, was shown in drawing 7 is made into an attention integrated case frame, now as subcategorization information Since NP [nom] showing being inevitably accompanied by the noun phrase used as a nominative case is generated as mentioned above the particle described by the attention integrated case frame of drawing 7 -- "-- it is -- " -- "-- " -- "-- alike -- " -- "-- ** -- " -- the case-marking particle which expresses a nominative case inside -- "-- " is recognized by the attention verb "it is conspicuous" as an indispensable case-marking particle. Furthermore, it sets to the attention integrated case frame of drawing 7 . a case-marking particle -- "-- as the attribute of " -- the case-marking particle -- "-- the noun which constitutes a nominative case with " Since it is what has the attribute [increase] or [thing] which cannot serve as an agent (agent) The attribute Theme showing the object as a superordinate concept of them is recognized. The attribute Theme as a subordinate concept The attribute Theme {thing/increase} showing an attribute [increase] and [thing] being included is recognized by the attention verb "it is conspicuous" as an attribute of an indispensable case-marking particle.

[0174] Then, it progresses to step S93 and the argument structure information generation section 24 recognizes the attribute to be an attention integrated case frame and the particle (suitably henceforth the particle of an option) by which the verb corresponding to an attention integrated case frame is accompanied if needed based on the subcategorization information.

[0175] That is, the argument structure information generation section 24 recognizes the thing excluding the indispensable case-marking particle recognized at step S92 from the particle described by the attention integrated case frame as a particle of an option. Furthermore, the argument structure information generation section 24 recognizes the attribute given to the particle recognized as a particle of an option as an attribute of the particle of an option in an attention integrated case frame.

[0176] When the integrated case frame about the verb "it is conspicuous" now shown in drawing 7 was made into an attention integrated case frame, as it followed, for example, it mentioned above an indispensable case-marking particle -- "-- the particle described by the attention integrated case frame of drawing 7 since it was " -- "-- it is -- " -- It is recognized as a particle of an option. "-- " -- "-- alike -- " -- "-- ** -- " -- from -- a case-marking particle -- "-- three particles except " -- "-- it is -- " -- "-- alike -- " -- "-- ** -- " -- as an attribute of the particle of the option further three particles described by the attention integrated case frame of drawing 7 -- "-- it is -- " -- "-- alike -- " -- "-- ** -- " -- each attribute Instrument, Locative, and Proposition is recognized.

[0177] And it progresses to step S94, and from the information recognized at steps S92 and S93, the argument structure information generation section 24 generates argument structure information, and outputs it to the auxiliary information generation section 25.

[0178] Namely, about the attention verb "it is conspicuous" corresponding to the attention integrated case frame shown in drawing 7 , as mentioned above, the argument structure information generation section 24 A set an indispensable case-marking particle -- "-- the attribute Theme {thing/increase} as " a list -- the case-marking particle and the set of an attribute of an option -- "-- it is -- " -- Instrument -- "-- alike -- " -- Locative -- and -- "-- ** -- ", when Proposition is obtained The argument structure information <ArgStr:Theme{thing/increase}-(Instrument)-(Locative)-(Proposition)> shown in drawing 8 is generated, and it outputs to the auxiliary information generation section 25.

[0179] Then, it progresses to step S95 and judges whether there is any integrated case frame which the argument structure information generation section 24 has not made an attention integrated case frame yet. In step S95, when judged with there being still an integrated case frame which has not been made into the attention integrated case frame, to step S91, return and the argument structure information generation section 24 still make one of the integrated case frames which have not been made into the

attention integrated case frame a new attention integrated case frame, and repeat the same processing hereafter.

[0180] Moreover, in step S95, when judged with there being still no integrated case frame which has not been made into the attention integrated case frame, subcategorization information generation processing is ended.

[0181] As mentioned above, according to the auxiliary information generation equipment of drawing 1, about much corpus data, a basic sentence is generated and the basic sentence to an unnecessary vocabulary is deleted from the morphological analysis result. Furthermore, about the verb in the basic sentence from which the unnecessary vocabulary was deleted, a case frame is generated and an integrated case frame is generated using the case frame about the same verb. And based on integrated each frame generated about each verb, the subcategorization information and argument structure information on the verb are generated, and it is outputted as auxiliary information. Therefore, when syntax analysis, a semantic analysis, etc. carry out natural language, it becomes possible to perform high syntax analysis and the high semantic analysis of precision by referring to the subcategorization information and argument structure information which are included in auxiliary information.

[0182] Next, drawing 16 shows the example of a configuration of the gestalt of other 1 operations of the natural-language-processing equipment which applied this invention.

[0183] This natural-language-processing equipment constitutes with voice the voice dialog system which performs a dialogue with a user.

[0184] That is, a microphone (microphone) 31 supplies the voice from a user to the A/D (Analog/Digital) converter 32 as a sound signal as an electrical signal. By carrying out A/D conversion of the sound signal of the analog from a microphone 31, A/D converter 32 is used as digital voice data, and is supplied to the speech recognition section 33. The speech recognition section 33 extracts feature vectors, such as MFCC (Mel Frequency Cepstrum Coefficient), by performing sonagraphy for the voice data from A/D converter 32 about a break and the voice data of each frame for every suitable frame. furthermore, the speech recognition section 33 -- the feature-vector sequence -- for example, HMM (Hidden Markov Model) -- matching processing is performed by law etc. and the voice inputted into the microphone 31 is recognized. The recognition result of the voice by the speech recognition section 33 is text data, and is supplied to the language-processing section 34.

[0185] By carrying out language processing of the speech recognition result from the speech recognition section 33, the language-processing section 34 generates the response sentence as a response (for example, a text) to the speech recognition result, and outputs it to the speech synthesis section 35, for example.

[0186] The composite tone corresponding to the response sentence from the language-processing section 34 is generated by performing for example, regulation speech synthesis processing, and the speech synthesis section 35 supplies it to the D/A (Digital/Analog) converter 36. D/A converter 36 is supplied to a loudspeaker 37 as a sound signal of an analog by carrying out D/A conversion of the digital composite tone data of the speech synthesis section 35. A loudspeaker 37 outputs the voice corresponding to the sound signal supplied from D/A converter 36, i.e., the composite tone corresponding to the response sentence generated in the language-processing section 34.

[0187] Next, in drawing 16, the language-processing section 34 consists of the morphological analysis section 41, the morphological analysis dictionary storage section 42, a syntax analyzer 43, the syntax-analysis dictionary storage section 44, a semantic analyzer 45, the auxiliary information database 46, the dialogue Management Department 47, a dialogue hysteresis database 48, and the response sentence generation section 49.

[0188] About the speech recognition result supplied from the speech recognition section 33, the morphological analysis section 41 performs morphological analysis, referring to the morphological analysis dictionary storage section 42, and supplies the morphological analysis result to a syntax analyzer 43. The morphological analysis dictionary storage section 42 has memorized the morphological analysis dictionary in which it refers to although the morphological analysis section 41 performs morphological analysis, for example, the reading, a functor attribute, a semantic attribute, etc. were

described about the morpheme.

[0189] Referring to a morphological analysis result, and the syntax-analysis dictionary storage section 44 and the auxiliary information database 46 from the morphological analysis section 41, a syntax analyzer 43 analyzes syntax of the speech recognition result of the speech recognition section 33, and supplies the syntax-analysis result to a semantic analyzer 45. The syntax-analysis dictionary storage section 44 was referred to for a syntax analyzer 43 to analyze syntax, for example, has memorized the syntax-analysis dictionary in which description about the dependency relation of a morpheme etc. is carried out.

[0190] Referring to the syntax-analysis result and the auxiliary information database 46 from a syntax analyzer 43, a semantic analyzer 45 performs the semantic analysis of the speech recognition result of the speech recognition section 33, and supplies the semantic-analysis result to the dialogue Management Department 47.

[0191] The auxiliary information database 46 has memorized the auxiliary information generated with the natural-language-processing equipment as auxiliary information generation equipment of drawing 1 about many verbs.

[0192] Referring to the semantic-analysis result and the dialogue hysteresis database 48 of the speech recognition result supplied from a semantic analyzer 45, the dialogue Management Department 47 understands the semantic content of the speech recognition result, generates the semantic content (suitably henceforth the content of a response) of the response sentence corresponding to the speech recognition result, and supplies the response sentence generation section 49.

[0193] The dialogue hysteresis database 48 memorizes the semantic content of a speech recognition result, and the content of a response which the dialogue Management Department 47 generated to the speech recognition result as dialogue hysteresis.

[0194] The response sentence generation section 49 generates the response sentence of the text corresponding to the content of a response from the dialogue Management Department 47, and supplies it to the speech synthesis section 35.

[0195] Next, with reference to the flow chart of drawing 17, the processing (interactive processing) which the voice dialog system of drawing 16 performs is explained.

[0196] A user's voice is inputted into a microphone 31, and further, if voice data is supplied to the speech recognition section 33 through A/D converter 32, in step S101, the speech recognition section 33 will carry out [voice / which was inputted into the microphone 31] speech recognition, will output the speech recognition result to the morphological analysis section 41 of the language-processing section 34, and will progress to step S102.

[0197] At step S102, the morphological analysis section 41 performs the morphological analysis by making the speech recognition result from the speech recognition section 33 into an input statement, supplies the morphological analysis result to a syntax analyzer 43, and progresses to step S103. At step S103, a syntax analyzer 43 retrieves the auxiliary information about the verb contained in the input statement from the auxiliary information database 46 by referring to the morphological analysis result of an input statement, and progresses to step S104 by it.

[0198] At step S104, based on the morphological analysis result from the morphological analysis section 41, an syntax-analysis dictionary, and the auxiliary information retrieved at step S103, a syntax analyzer 43 analyzes the syntax of the speech recognition result as an input statement, and supplies the syntax-analysis result to a semantic analyzer 45. Furthermore, at step S104, a semantic analyzer 45 performs a semantic analysis based on the syntax-analysis result of the speech recognition result as an input statement supplied from a syntax analyzer 43, and it progresses to step S105.

[0199] At step S105, the indispensable noun lacks in the verb contained in whether an anaphor exists in an input statement, and its input statement, or (zero anaphor) it is judged whether the pronoun is substituted for the indispensable noun.

[0200] In addition, it can be recognized in syntax analysis by the syntax analyzer 43 whether an anaphor exists in an input statement.

[0201] That is, according to the subcategorization information included in the auxiliary information

about a verb "it is conspicuous" shown in drawing 8, for example, it turns out that a verb "it is conspicuous" is inevitably accompanied by the noun phrase used as a nominative case. Therefore, if the original form has not followed on it the noun phrase from which the verb "it is conspicuous" serves as a nominative case when the "conspicuous" verb is contained by the input statement, the noun phrase indispensable about the verb "it is conspicuous" is missing from the auxiliary information about a verb "it is conspicuous", namely, a syntax analyzer 43 can recognize that a zero anaphor exists. In addition, the existence of an anaphor can be recognized also by the function of SACHURESHON (saturation) in frameworks, such as HPSG.

[0202] In step S105, when judged with an anaphor not existing in an input statement, a semantic analyzer 45 supplies the semantic-analysis result of an input statement to the dialogue Management Department 47, skips step S106 thru/or step S110, and progresses to step S111.

[0203] Moreover, in step S105, when judged with an anaphor existing in an input statement, a semantic analyzer 45 recognizes the attribute of an anaphor by referring to the auxiliary information database 46 by progressing to step S106.

[0204] That is, at step S106, a semantic analyzer 45 recognizes the attribute of the noun by which the verb contained in an input statement should be accompanied inevitably from the subcategorization information on the auxiliary information retrieved at step S103, and argument structure information. and the attribute of the noun on which the verb contained in the input statement should follow a semantic analyzer 45 inevitably -- the attribute of the noun which lacks in the speech recognition result, or the noun for which the pronoun is substituted is recognized inside.

[0205] Then, the semantic Management Department 45 judges whether the noun of the same attribute as the attribute of the anaphor recognized at step S106 exists in the dialogue hysteresis of the dialogue hysteresis database 48 by asking the dialogue Management Department 47 by progressing to step S107.

[0206] In addition, at step S107, it judges whether the noun of the same attribute as the attribute of an anaphor exists for the dialogue hysteresis of the range before 1 thru/or 4 utterance, for example according to the heuristics (Minimal Distance Principle) that the distance of the antecedent and correspondence house which are advocated by J.Huang, "Logical Relations in Chinese and Theory of Grammar", MIT PhD.Thesis, and 1982 is minimal.

[0207] In step S107, when it judges that the noun of the same attribute as the attribute of an anaphor does not exist in the dialogue hysteresis of the dialogue hysteresis database 48, it progresses to step S108 and the dialogue Management Department 47 performs inquiry processing which asks the content of the anaphor to a user.

[0208] That is, the dialogue Management Department 47 makes the response sentence generation section 49 generate the message (suitably henceforth an inquiry message) which asks the content of the anaphor, and makes it output by composite tone from a loudspeaker 37 through the speech synthesis section 35 and D/A converter 36.

[0209] And if a user performs the utterance explaining the content of the anaphor corresponding to an inquiry message, the voice will be supplied to a semantic analyzer 45 through a microphone 31, A/D converter 32, the speech recognition section 33, the morphological analysis section 41, and a syntax analyzer 43.

[0210] A semantic analyzer 45 is carried out in this way, from a syntax analyzer 43, it waits to supply the syntax-analysis result about the voice of the user explaining the content of the anaphor, progresses to S109 from step S108, is based on the syntax-analysis result, recognizes and determines the antecedent of an anaphor, and progresses to step S110.

[0211] On the other hand, when it judges that the noun of the same attribute as the attribute of an anaphor exists in the dialogue hysteresis of the dialogue hysteresis database 48 in step S107, it progresses to step S109, and a semantic analyzer 43 determines the noun of the same attribute as the anaphor which exists in the dialogue hysteresis as an antecedent of the anaphor, and progresses to step S110.

[0212] At step S110, as that in which the antecedent determined at step S109 exists instead of the anaphor in an input statement, a syntax analyzer 43 analyzes syntax, and further, a semantic analyzer 45

performs a semantic analysis and supplies the semantic-analysis result to the dialogue Management Department 47 about the input statement.

[0213] If the semantic-analysis result of an input statement is received from a semantic analyzer 45, the dialogue Management Department 47 progresses to step S111, based on the semantic-analysis result, it will understand the semantics of an input statement, will generate the content (the content of a response) of the response sentence as a response corresponding to the input statement, and will progress to step S112. At step S112, the dialogue Management Department 47 supplies the content of a response to the response sentence generation section 49, and progresses to step S113 while it supplies the set of the semantic content of an input statement, and the semantic content (the content of a response) of the generated response sentence to the dialogue hysteresis database 48 and makes it memorize as dialogue hysteresis.

[0214] At step S113, the response sentence generation section 49 generates the response sentence which makes the content of a response from the dialogue Management Department 47 the semantic content, and supplies it to the speech synthesis section 35. Furthermore, at step S112, the speech synthesis section 35 generates the composite tone corresponding to the response sentence from the response sentence generation section 49, makes it output from a loudspeaker 37 through D/A converter 36, and closes interactive processing.

[0215] In addition, in the above interactive processing, although it was made to ask a user when the antecedent of an anaphor was not able to be determined from dialogue hysteresis and was not able to be determined from dialogue hysteresis in principle, it determines from dialogue hysteresis and the antecedent of an anaphor can also be made not to ask a user, either.

[0216] However, it is necessary to except the case which has the antecedent of an anaphor in the interior of the same sentence in that case, and the case for which an understanding (deictic use) accompanied by directions or vision is required.

[0217] Here, if the case which has the antecedent of an anaphor in the interior of the same sentence expresses an anaphor as pro, the sentence "the man by whom the paper written pro was commended" corresponds, for example. The anaphor pro in this sentence is pointing to the man (man to whom the idea of the written paper was carried out) who has said in this sentence, and the "man" who becomes the antecedent of an anaphor is in the interior of the same sentence. Thus, the problem of an anaphor in case the antecedent of an anaphor is in the interior of the same sentence is for example, the Iwanami lecture. Science 6 "generative grammar" Iwanami Shoten of language It is solvable with a bounding theory (binding theory) which will exist in 1997 etc.

[0218] Moreover, a case [need / with directions or rating in the antecedent of an anaphor / to be understood] is the case where pointed at the cop on a desk and "Gather it" is said.

[0219] In addition, also about which case, if a user is asked, it is possible to determine the antecedent of an anaphor.

[0220] According to interactive processing of drawing 17 , the antecedent of an anaphor is determined as follows, for example.

[0221] That is, for example, a voice dialog system outputs composite tone "Mr. A ate the eel on the day of the dog days.", and presupposes now that the user spoke saying, "Has Mr. B already eaten?" to it.

[0222] in this case -- in order for a voice dialog system to understand a user's utterance correctly -- a user's utterance "has Mr. B already eaten?" -- "an eel" -- compensating -- "Mr. B -- already -- "an eel" -- it ate -- " -- ** -- it is necessary to carry out

[0223] Then, refer to the auxiliary information about the verb (original form) "it eats" contained in a user's utterance "has Mr. B already eaten?" for a voice dialog system.

[0224] Now, the auxiliary information about a verb "it eats" presupposes that it was a thing as shown in drawing 18 .

[0225] Here, the 1st line (from a top to the 1st line) of the auxiliary information about the verb "it eats" in drawing 18 expresses a verbal header "it eats", reading "TABERU", and classification information "a transitive verb." Moreover, it means being inevitably accompanied by the noun phrase (NP [acc]) to which the subcategorization information on the 2nd line <SUBCAT:NP[nom]-NP[acc]> expresses the

noun phrase (NP [nom]) to which a verb "it eats" expresses a nominative case (nominative), and an accusative (accusative). Furthermore, the argument structure information on the 3rd line <ArgStr:Agent-Theme{food}-(Instrument)-(Locative)> The noun phrase NP showing the nominative case of subcategorization information [nom] is a thing used as a verbal "it eats" agent (Agent), Noun phrase NP= [acc] showing the accusative of subcategorization information is a thing used as a verbal "it eats" object (Theme), The object's (Theme's) being food {food} and a verb "it eats" mean that the particle to which an attribute is expressed with Instrument or Locative can be taken if needed.

[0226] In addition, attributes Instrument and Locative express instruments (for example, "knife" etc.) and a location "at for example, restaurant", respectively, as mentioned above.

[0227] About a user's utterance "has Mr. B already eaten?", it is a noun phrase showing the physique and by referring to the auxiliary information on drawing 18 shows what the thing showing the food used as the object to eat is [the thing] missing (a zero anaphor exists).

[0228] On the other hand, in now, the voice dialog system is outputting "Mr. A ate the eel on the day of the dog days" just before a user's utterance "has Mr. B already eaten?", and the "eel" of this output is a noun phrase showing the physique, and expresses the food used as the object to eat.

[0229] Therefore, in this case, a voice dialog system is a noun phrase showing an accusative which lacks in a user's utterance "has Mr. B already eaten?" by referring to dialogue hysteresis, and the thing showing the food used as the object to eat can recognize that it is an "eel." That is, it is determined that the antecedent of the zero anaphor which exists in a user's utterance "has Mr. B already eaten?" in this case is an "eel."

[0230] consequently, the antecedent "an eel" which determined the voice dialog system as a user's utterance "has Mr. B already eaten?" -- compensating -- "-- Mr. B -- already -- "an eel" -- it ate -- " -- ** -- he can carry out and can understand the semantic content correctly.

[0231] in addition, when the thing showing the food used as the object which is a noun phrase showing an accusative and is eaten does not exist in dialogue hysteresis as the message which, as for a voice dialog system, the food asks that it is what it is -- for example, -- "-- Mr. B ate what -- ?" etc. is generated and outputted, it waits for answerback of the user to the message, and the antecedent (in now, it is an "eel") of a zero anaphor is determined.

[0232] Moreover, in the above-mentioned case, were aimed at the time of being "whether Mr. B has already eaten" at which a user's utterance has a zero anaphor, but According to interactive processing of drawing 17, a user's utterance has the anaphor which is not a zero anaphor, for example, as well as the case in a zero anaphor when it is "whether whether Mr. B to already have eaten it (that)", the antecedent of an anaphor "it (be)" can be determined.

[0233] as mentioned above, in the voice dialog system of drawing 16 By referring to auxiliary information including the subcategorization information and argument structure information on verbal Since the antecedent to which the anaphor points is determined based on the attribute of the anaphor and it was made to perform syntax analysis or the semantic analysis of an input statement after having recognized the attribute of the anaphor which exists in an input statement High syntax analysis and the high semantic analysis of precision become possible, and further this becomes possible to understand the semantics of an input statement to accuracy.

[0234] In addition, although it was made to include classification information in auxiliary information with the gestalt of this operation, auxiliary information can be constituted, without including classification information. However, it can be said that classification information is indirectly included in it since classification information can be acquired from subcategorization information to auxiliary information even when classification information is not included clearly.

[0235] Next, hardware can also perform a series of processings mentioned above, and software can also perform. When software performs a series of processings, the program which constitutes the software is installed in a general-purpose computer etc.

[0236] Then, drawing 19 shows the example of a configuration of the gestalt of 1 operation of the computer by which the program which performs a series of processings mentioned above is installed.

[0237] A program is recordable on the hard disk 105 and ROM103 as a record medium which are built

in the computer beforehand.

[0238] Or a program is permanently [temporarily or] storable in the removable record media 111, such as a flexible disk, CD-ROM (Compact DiscRead Only Memory), MO (Magneto Optical) disk, DVD (Digital Versatile Disc), a magnetic disk, and semiconductor memory, again (record). Such a removable record medium 111 can be offered as the so-called software package.

[0239] In addition, it installs in a computer from the removable record medium 111 which was mentioned above, and also from a download site, through the satellite for digital satellite broadcasting services, it transmits to a computer on radio, or a program is transmitted to a computer with a cable through networks, such as LAN (Local Area Network) and the Internet, and by computer, it can receive in the communications department 108 and it can install the program transmitted by making it such on the hard disk 105 to build in.

[0240] The computer contains CPU (Central Processing Unit)102. The input/output interface 110 is connected to CPU102 through the bus 101, and the input section 107 from which CPU102 is constituted from a keyboard, a mouse, a microphone, etc. by the user through an input/output interface 110 will perform the program stored in ROM (Read Only Memory)103 according to it, if a command is inputted by [, such as actuation,] being carried out. Or it is transmitted from the program and satellite with which CPU102 is stored in the hard disk 105 again, or a network, and the program which was received in the communications department 108 and installed on the hard disk 105, or the program which reading appearance was carried out from the removable record medium 111 with which the drive 109 was equipped, and was installed on the hard disk 105 is loaded to RAM (Random Access Memory)104, and is performed. Thereby, CPU102 performs processing performed by the configuration of the block diagram according to the flow chart mentioned above processed or mentioned above. and the output from the output section 106 by which CPU102 is constituted from LCD (Liquid CryStal Display), a loudspeaker, etc. through an input/output interface 110 in the processing result if needed or the transmission from the communications department 108 -- record etc. is further carried out to a hard disk 105.

[0241] It is not necessary to necessarily process the processing step which describes the program for making various kinds of processings perform to a computer in this description here to time series in accordance with the sequence indicated as a flow chart, and it is a juxtaposition thing also including the processing (for example, parallel processing or processing by the object) performed according to an individual.

[0242] Moreover, a program may be processed by the computer of 1 and distributed processing may be carried out by two or more computers. Furthermore, a program may be transmitted to a distant computer and may be executed.

[0243] In addition, auxiliary information can be used by the system which performs natural language processing of an others and text epitome, a translation, and others shown in drawing 16 . [dialog system / voice] Moreover, auxiliary information is stored in the independent auxiliary information database 46 as shown in drawing 16 , and also it is possible to make it memorize in the form unified to REKISHIKON (dictionary) (for example, the morphological analysis dictionary of the morphological analysis dictionary storage section 42 of drawing 17 , the syntax-analysis dictionary of the syntax-analysis dictionary storage section 44, etc.) used by the system.

[0244] Moreover, this invention is applicable also to natural language other than Japanese.

[0245]

[Effect of the Invention] According to the program, the basic sentence which is the unit made applicable [of a case frame] to generation is generated by the 1st natural-language-processing equipment of this invention and the natural-language-processing approach, and the list from the morphological analysis result of corpus data, and an unnecessary vocabulary is deleted from the basic sentence by generation of a case frame. Furthermore, about the verb in the basic sentence from which the unnecessary vocabulary was deleted, a case frame is generated, the subcategorization information and argument structure information on the verb are generated based on the case frame about the same verb, and it is outputted as auxiliary information. Therefore, high syntax analysis, a high semantic analysis, etc. of precision

become possible by referring to the auxiliary information.

[0246] According to the program, in the 2nd natural-language-processing equipment of this invention and the natural-language-processing approach, and a list From an auxiliary information storage means by which the auxiliary information which consists of the subcategorization information and argument structure information on verbal is memorized at least While the auxiliary information about the verb contained in an input statement is retrieved, the attribute of the anaphor to which it is judged whether an anaphor exists in an input statement, and it exists in an input statement is recognized based on the auxiliary information about the verb contained in the input statement. And the antecedent to which an anaphor points is determined based on the attribute of an anaphor, and syntax analysis or the semantic analysis of an input statement is performed using the antecedent. Therefore, high syntax analysis, a high semantic analysis, etc. of precision become possible, and further this becomes possible to understand the semantics of an input statement to accuracy.

[Translation done.]

* NOTICES *

JPO and INPIT are not responsible for any damages caused by the use of this translation.

- 1.This document has been translated by computer. So the translation may not reflect the original precisely.
- 2.**** shows the word which can not be translated.
- 3.In the drawings, any words are not translated.

DESCRIPTION OF DRAWINGS

[Brief Description of the Drawings]

[Drawing 1] It is the block diagram showing the example of a configuration of the gestalt of 1 operation of the natural-language-processing equipment which applied this invention.

[Drawing 2] It is drawing showing a morphological analysis result.

[Drawing 3] It is drawing explaining the vocabulary (unnecessary vocabulary) deleted from a basic sentence.

[Drawing 4] It is drawing showing the morphological analysis result from which the unnecessary vocabulary was deleted.

[Drawing 5] It is drawing explaining a verbal criteria form.

[Drawing 6] It is drawing showing a case frame.

[Drawing 7] It is drawing showing an integrated case frame.

[Drawing 8] It is drawing showing auxiliary information.

[Drawing 9] It is a flow chart explaining auxiliary information generation processing.

[Drawing 10] It is a flow chart explaining basic sentence pattern extract processing.

[Drawing 11] It is a flow chart explaining unnecessary lexical deletion.

[Drawing 12] It is a flow chart explaining case frame generation processing.

[Drawing 13] It is a flow chart explaining verb classification processing.

[Drawing 14] It is a flow chart explaining subcategorization information generation processing.

[Drawing 15] It is a flow chart explaining argument structure information generation processing.

[Drawing 16] It is the block diagram showing the example of a configuration of the gestalt of other 1 operations of the natural-language-processing equipment which applied this invention.

[Drawing 17] It is a flow chart explaining interactive processing.

[Drawing 18] It is drawing showing auxiliary information.

[Drawing 19] It is the block diagram showing the example of a configuration of the gestalt of 1 operation of the computer which applied this invention.

[Description of Notations]

1 Corpus Database 2 Pretreatment Section, 3 Case frame database 4 case frame processing section 5 An auxiliary information database, 11 Morphological analysis section 12 Basic sentence pattern extract section 13 Cutout 14 Case frame generation section 21 Case frame integrated section 22 Verb classification section 23 The subcategorization information generation section, 24 Argument structure information generation section 25 The auxiliary information generation section, 31 microphone 32 An A/D converter, 33 Speech recognition section 34 The language-processing section and 35 speech synthesis section 36 A D/A converter, 38 Loudspeaker 41 The morphological analysis section, 42 Morphological analysis dictionary storage section 43 syntax analyzer 44 The syntax-analysis dictionary storage section, 45 Semantic analyzer 46 An auxiliary information database, 47 dialogue Management Department 48 A dialogue hysteresis database, 49 Response sentence generation section 101 A bus, 102CPU 103 ROM, 104 RAM 105 Hard disk 106 Output section 107 Input section 108 Communications department 109 drives 110 Input/output interface 111 Removable record medium

[Translation done.]

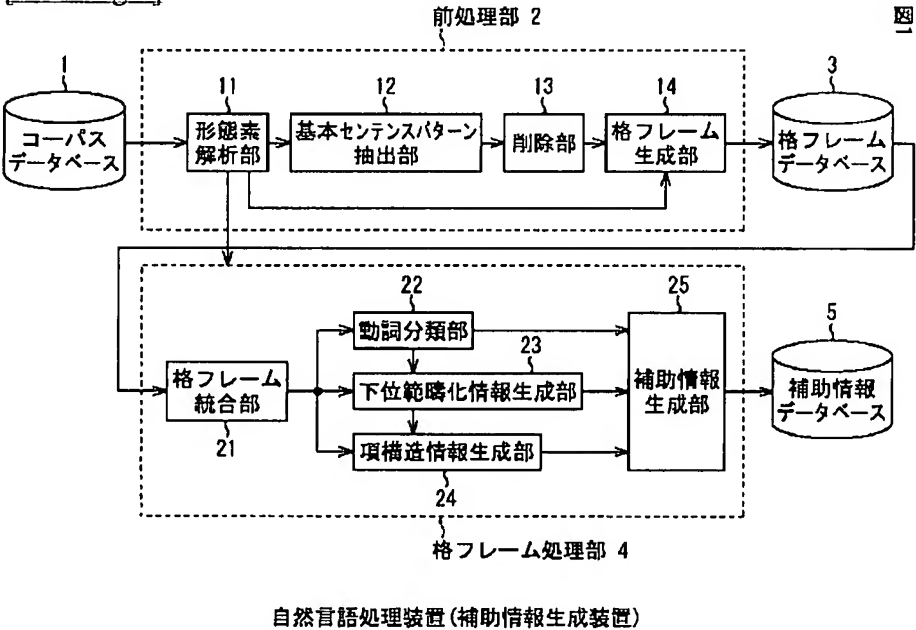
* NOTICES *

JPO and INPIT are not responsible for any damages caused by the use of this translation.

- 1.This document has been translated by computer. So the translation may not reflect the original precisely.
- 2.**** shows the word which can not be translated.
- 3.In the drawings, any words are not translated.

DRAWINGS

[Drawing 1]



[Drawing 2]

| 見出し | 読み | シソーラス情報(構文属性、意味属性、動詞の原型) |
|------|-------------|--|
| 特に | トクニ | [[CAT Adverb][VAL 特に]] |
| 県内果実 | ケンナイカジツ | [[CAT Noun][ol Compound=CN+CN][Sem food][VAL 県内果実]] |
| が | ガ | [[CAT Case][cl abstract][fx nominative][VAL が]] |
| 数量 | スーリョー | [[CAT Noun][cl CNoun][Sem amount][VAL 数量]] |
| で | デ | [[CAT Case][cl lexical][fx instrument][VAL で]] |
| 一八%増 | イチハチパーセントゾー | [[CAT Noun][cl Compound=Num+Classifier+suf][Sem increase][VAL一八%増]] |
| 、 | キンガク | [[CAT Punctuation][cl comma][VAL 、]] |
| 金額 | デ | [[CAT Noun][cl CNoun][Sem money][VAL 金額]] |
| で | サンヨンパーセントゾー | [[CAT Case][ol lexical][fx instrument][VAL で]] |
| 三四%増 | | [[CAT Noun][cl Compound=Num+Classifier+suf][Sem increase][VAL三四%増]] |
| と | ト | [[CAT Complementizer][ol proposition][VAL と]] |
| 伸び | ノビ | [[CAT Noun][cl CNoun][Sem increase][VAL 伸び]] |
| が | ガ | [[CAT Case][cl abstract][fx nominative][VAL が]] |
| 目立った | メダッタ | [[CAT Verb][cl active][fm finite][Conj(c12)(Stem 目立つ)(fm aff-past)(Polarity aff)(Ts past)] |
| . | . | [[Style(cl plain)(fm zero)][VAL 目立った]] |
| | | [[CAT Punctuation][cl period][VAL 。]] |

形態素解析結果

[Drawing 3]

図3

- (A) 副詞
[[CAT Adverb]].
- (B) 名詞+「の」(例:夏場の)
[[CAT Noun]...]
[[CAT Case][cl abstract][fx genitive][VAL の]]
- (C) 名詞+助詞+「の」(例:日本での)
[[CAT Noun]...]
[[CAT Case]...]
[[CAT Case][cl abstract][fx genitive][VAL の]]
- (D) 形容詞
[[CAT Adjective][cl stative]...]
- (E) 名詞(形容動詞語幹)+「な」(例:決定的な)
[[CAT Noun][cl AdjNoun]...]
[[CAT Verb][cl copula]...[VAL な]].
- (F) 名詞+後置詞(例:工場に対する)
[[CAT Noun]...]
[[CAT Postposition]...]
- (G) 括弧内の文書
[[CAT Punctuation][cl L-]]
[[CAT Punctuation][cl R-]]
- (H) 括弧内の文書+「の」
[[CAT Punctuation][cl L-]]
[[CAT Punctuation][cl R-]]
[[CAT Case][cl abstract][fx genitive][VAL の]]

削除される語彙

[Drawing 4]

図4

| 見出し | 読み | シソーラス情報 |
|------|-------------|---|
| 県内果実 | ケンナйкаジツ | [[CAT Noun][cl Compound=CN+CN][Sem food][VAL 県内果実]] |
| が | ガ | [[CAT Case][cl abstract][fx nominative][VAL が]] |
| 数量 | スーリョー | [[CAT Noun][cl CNoun][Sem amount][VAL 数量]] |
| で | デ | [[CAT Case][cl lexical][fx instrument][VAL で]] |
| 一八%増 | イチハチパーセントゾー | [[CAT Noun][cl Compound=Num+Classifier+suf][Sem increase][VAL 一八%増]] |
| 、 | キンガク | [[CAT Punctuation][cl comma][VAL 、]] |
| 金額 | デ | [[CAT Noun][cl CNoun][Sem money][VAL 金額]] |
| で | サンヨンパーセントゾー | [[CAT Case][cl lexical][fx instrument][VAL で]] |
| 三四%増 | | [[CAT Noun][cl Compound=Num+Classifier+suf][Sem increase][VAL 三四%増]] |
| と | ト | [[CAT Complementizer][cl proposition][VAL と]] |
| 伸び | ノビ | [[CAT Noun][cl CNoun][Sem increase][VAL 伸び]] |
| か | カ | [[CAT Case][cl abstract][fx nominative][VAL が]] |
| 目立った | メダッタ | [[CAT Verb][cl active][fm finite][Conj(c12)(Stem 目立つ)(fm aff-past)(Polarity aff)(Ts past)][Style(cl plain)(fm zero)][VAL 目立った]] |

削除処理後の基本センテンス

[Drawing 7]

図7

目立つ メダツ subcat: で[Instrument]:
 が[increase], [thing]:
 に[Locative]:
 と[Proposition]:]

統合格フレーム

[Drawing 5]

図5

| 見出し | 読み | シソーラス情報 |
|------|------|---|
| 目立つ | メダツ | [[CAT Verb][cl active][fm finite][Conj (cl2) (fm aff-non-past) (Stem 目立つ) (Polarity aff) (Ts non-past)]...] |
| 目立った | メダッタ | [[CAT Verb][cl active][fm finite][Conj (cl2) (fm aff-past) (Stem 目立つ) (Polarity aff) (Ts past)]...] |

(A)

| 見出し | 読み | シソーラス情報 |
|----------|--------|---|
| 適用 する | テキヨースル | [[CAT Noun][cl Ynoun]...[VAL 適用]] [[CAT Verb][cl active][fm finite]... (Stem する) (fm aff-non-past) ...[VAL する]] |

(B)

| 見出し | 読み | シソーラス情報 |
|------------|------------|--|
| 見込んで いる | ミコンデ イル | [[CAT Verb]...[fm infinite]... (Stem 見込む) (fm pres-participle) ...[VAL 見込んで]] [[CAT Verb][cl active][fm finite]... (Stem いる) ...[VAL いる]] |

(C)

| 見出し | 読み | シソーラス情報 |
|----------------|------------------|---|
| 展開 して いる | テンカイ シテ イル | [[CAT Noun][cl Ynoun]...[VAL 展開]] [[CAT Verb]...[fm finite]... (Stem する) (fm pres-participle) ...[VAL して]] [[CAT Verb]...[fm finite]... (Stem いる) ... [VAL いる]] |

(D)

動詞の基準形

[Drawing 6]

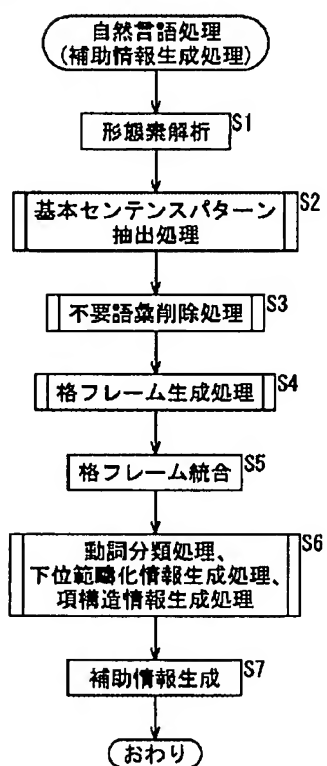
図
9

[目立つ C_FRAME: で[instrument], が[increase]]
 [目立つ C_FRAME: が[thing]]
 [目立つ C_FRAME: と[proposition], が[thing]]
 [目立つ C_FRAME: で[instrument], に[locative], が[increase]]

格フレーム

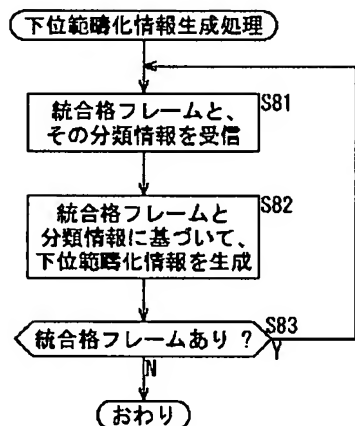
[Drawing 9]

図9



[Drawing 14]

図14



[Drawing 8]

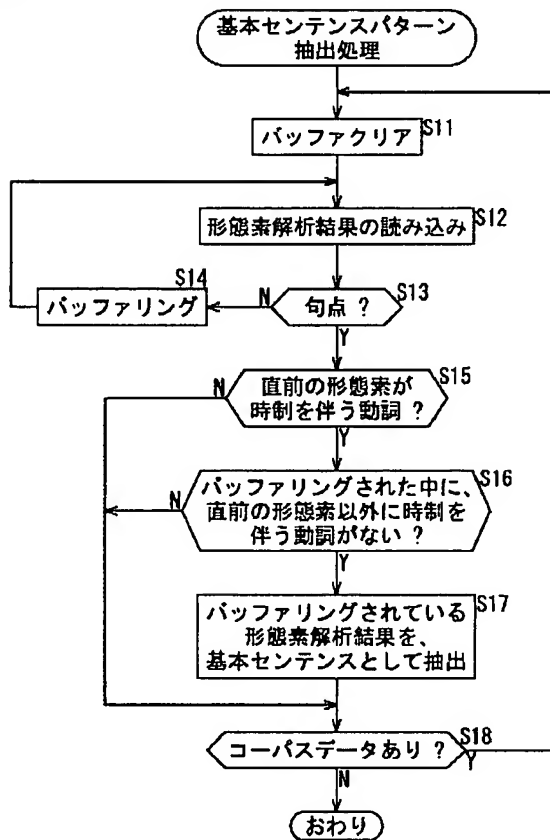
図
8

{目立つ メダツ 能動動詞
 下位範疇化情報: <SUBCAT:NP[nom]>
 項構造情報: <ArgStr:Theme[thing/increase]
 -(Instrument)-(Locative)-(Proposition)>
 }

補助情報

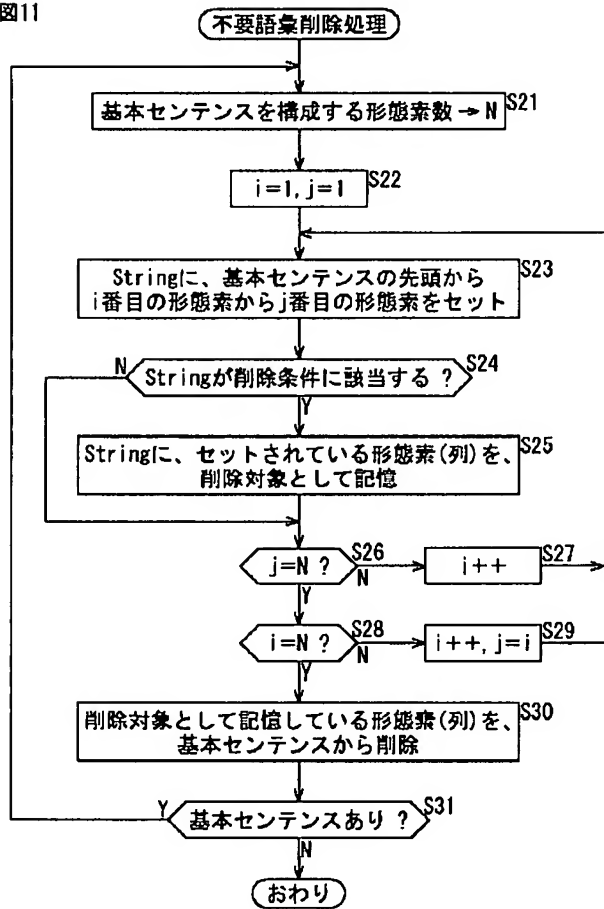
[Drawing 10]

図10



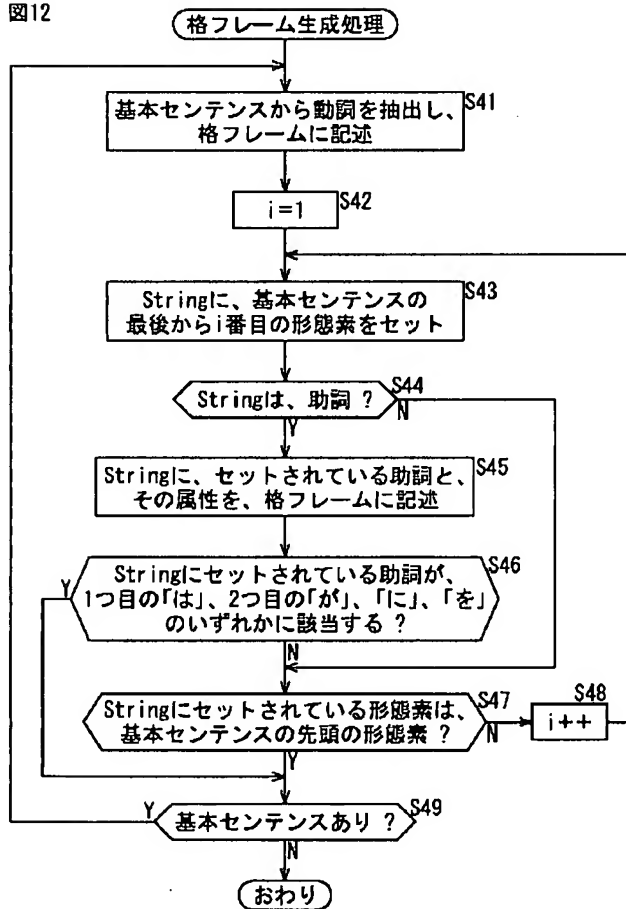
[Drawing 11]

図11



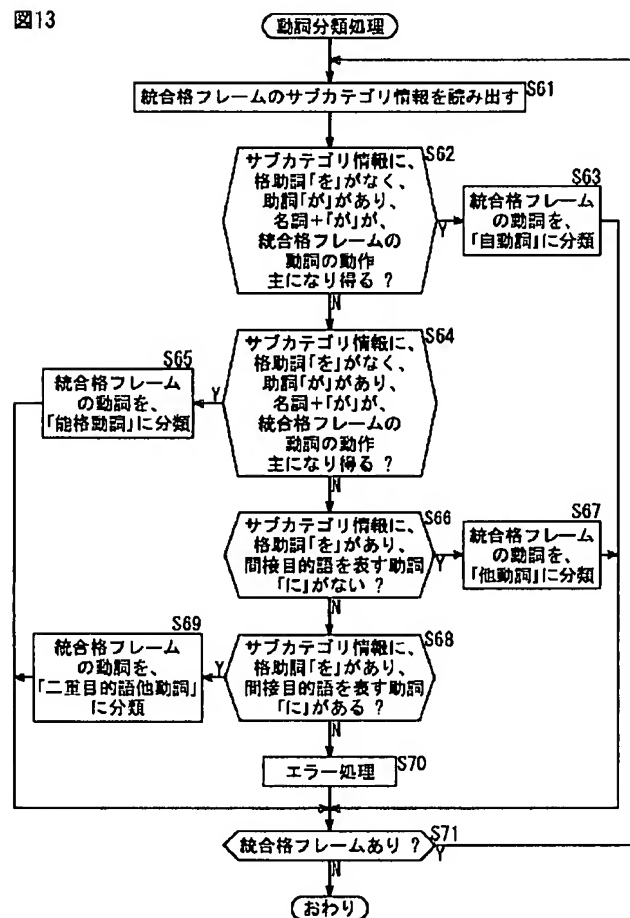
[Drawing 12]

図12



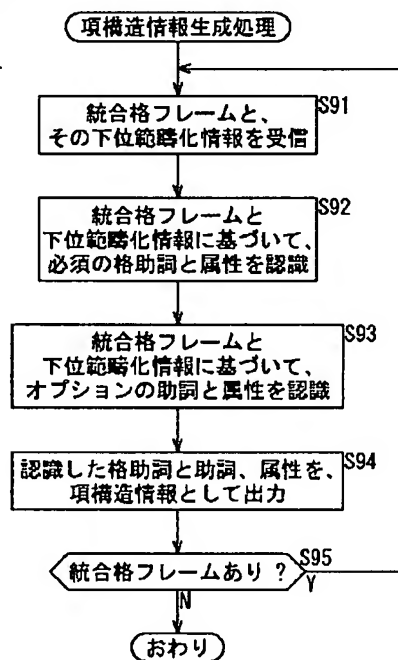
[Drawing 13]

図13

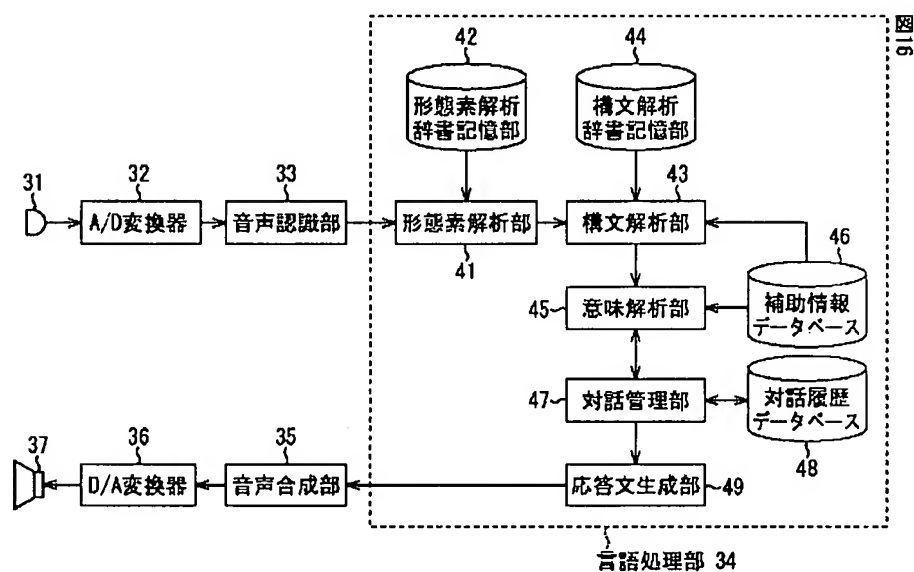


[Drawing 15]

図15



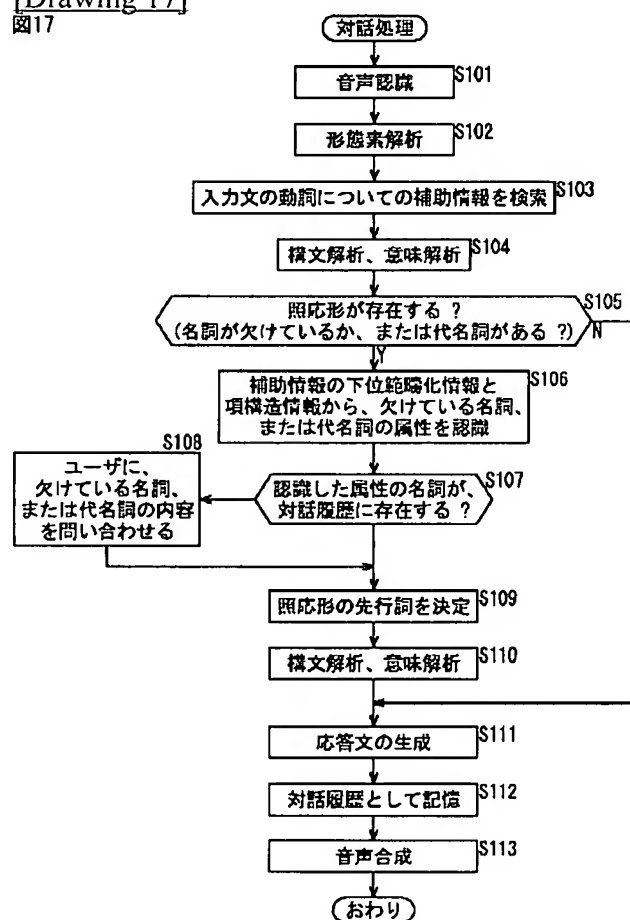
[Drawing 16]



自然言語処理装置 (音声対話システム)

[Drawing 17]

図17



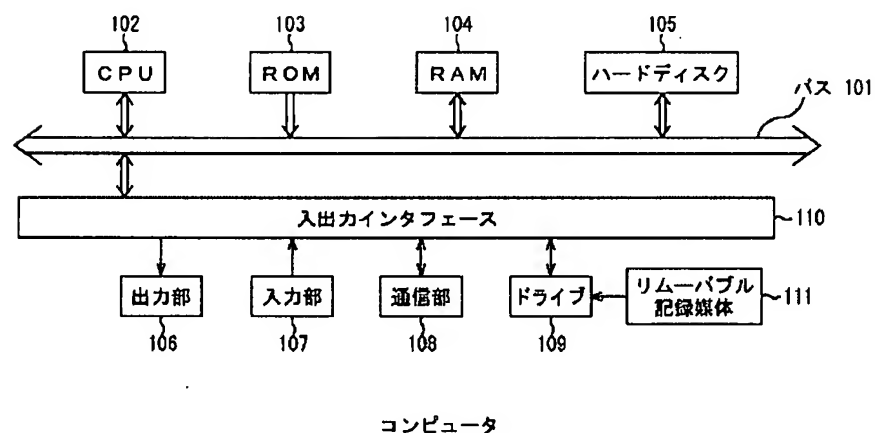
[Drawing 18]

図
8

[食べる タベル 他動詞
 下位範疇化情報: <SUBCAT:NP[nom]-NP[acc]>
 項構造情報: <ArgStr:Agent-Theme {food}
 -(Instrument)-(Locative)>
]

「食べる」についての補助情報

[Drawing 19]

図
19

[Translation done.]

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号
特開2003-30184
(P2003-30184A)

(43) 公開日 平成15年1月31日 (2003.1.31)

(51) Int.Cl.⁷
G 0 6 F 17/27

識別記号

F I
G 0 6 F 17/27

データベース* (参考)
M 5 B 0 9 1
E

審査請求 未請求 請求項の数17 OL (全 27 頁)

(21) 出願番号 特願2001-217619(P2001-217619)

(22) 出願日 平成13年7月18日 (2001.7.18)

(71) 出願人 000002185

ソニー株式会社

東京都品川区北品川6丁目7番35号

(72) 発明者 田島 和彦

東京都品川区北品川6丁目7番35号 ソニ
ー株式会社内

(72) 発明者 横田 重昭

東京都品川区北品川6丁目7番35号 ソニ
ー株式会社内

(74) 代理人 100082131

弁理士 稲本 義雄

最終頁に続く

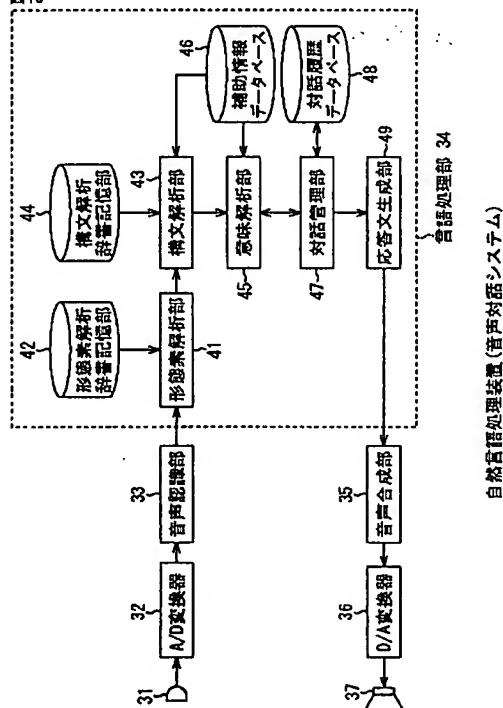
(54) 【発明の名称】 自然言語処理装置および自然言語処理方法、並びにプログラムおよび記録媒体

(57) 【要約】

【課題】 精度の高い構文解析や意味解析を行い、入力文の意味を正確に理解する。

【解決手段】 意味解析部45は、動詞の下位範疇化情報と項構造情報からなる、多量のコーパスデータを用いて生成された補助情報を記憶している補助情報データベース46から、入力文に含まれる動詞についての補助情報を検索し、入力文中に存在する照応形の属性を、その入力文に含まれる動詞についての補助情報に基づいて認識する。そして、意味解析部45は、照応形の属性に基づいて、照応形が指し示す先行詞を決定し、その先行詞を用いて、入力文の意味解析を行う。

図16



【特許請求の範囲】

【請求項1】 自然言語の解析を補助する補助情報を、コーパスデータから求める自然言語処理装置であって、前記コーパスデータを形態素解析する形態素解析手段と、

前記コーパスデータの形態素解析結果から、格フレームの生成対象とする単位である基本センテンスを生成する基本センテンス生成手段と、

前記基本センテンスから、格フレームの生成に不要な語彙を削除する不要語彙削除手段と、

前記不要語彙が削除された基本センテンスにおける動詞について、格フレームを生成する格フレーム生成手段と、

同一の動詞についての格フレームに基づいて、その動詞の下位範疇化情報と項構造情報を生成し、前記補助情報として出力する補助情報生成手段とを備えることを特徴とする自然言語処理装置。

【請求項2】 前記補助情報生成手段は、同一の動詞についての格フレームに基づいて、その動詞が、自動詞、他動詞、能格動詞、または二重目的語他動詞のうちのいずれに分類されるものであるかを表す分類情報を生成し、前記分類情報に基づいて、前記下位範疇化情報を生成することを特徴とする請求項1に記載の自然言語処理装置。

【請求項3】 前記不要語彙削除手段は、副詞、名詞と「の」からなる語彙、名詞と助詞と「の」からなる語彙、形容詞、名詞と「な」からなる語彙、名詞と後置詞からなる語彙、括弧で囲まれた部分、または括弧で囲まれた部分と「の」からなる語彙を、前記基本センテンスから削除することを特徴とする請求項1に記載の自然言語処理装置。

【請求項4】 前記補助情報生成手段は、同一の動詞についての格フレームの格助詞に基づいて、前記下位範疇化情報を生成することを特徴とする請求項1に記載の自然言語処理装置。

【請求項5】 前記補助情報生成手段は、同一の動詞についての格フレームすべての助詞に基づいて、前記項構造情報を生成することを特徴とする請求項1に記載の自然言語処理装置。

【請求項6】 前記コーパスデータは、日本語のデータであることを特徴とする請求項1に記載の自然言語処理装置。

【請求項7】 自然言語の解析を補助する補助情報を、コーパスデータから求める自然言語処理方法であって、前記コーパスデータを形態素解析する形態素解析ステップと、

前記コーパスデータの形態素解析結果から、格フレームの生成対象とする単位である基本センテンスを生成する基本センテンス生成ステップと、

前記基本センテンスから、格フレームの生成に不要な語

彙を削除する不要語彙削除ステップと、

前記不要語彙が削除された基本センテンスにおける動詞について、格フレームを生成する格フレーム生成ステップと、

同一の動詞についての格フレームに基づいて、その動詞の下位範疇化情報と項構造情報を生成し、前記補助情報として出力する補助情報生成ステップとを備えることを特徴とする自然言語処理方法。

【請求項8】 自然言語の解析を補助する補助情報を、コーパスデータから求める自然言語処理を、コンピュータに行わせるプログラムであって、

前記コーパスデータを形態素解析する形態素解析ステップと、

前記コーパスデータの形態素解析結果から、格フレームの生成対象とする単位である基本センテンスを生成する基本センテンス生成ステップと、

前記基本センテンスから、格フレームの生成に不要な語彙を削除する不要語彙削除ステップと、

前記不要語彙が削除された基本センテンスにおける動詞について、格フレームを生成する格フレーム生成ステップと、

同一の動詞についての格フレームに基づいて、その動詞の下位範疇化情報と項構造情報を生成し、前記補助情報として出力する補助情報生成ステップとを備えることを特徴とするプログラム。

【請求項9】 自然言語の解析を補助する補助情報を、コーパスデータから求める自然言語処理を、コンピュータに行わせるプログラムが記録されている記録媒体であって、

前記コーパスデータを形態素解析する形態素解析ステップと、

前記コーパスデータの形態素解析結果から、格フレームの生成対象とする単位である基本センテンスを生成する基本センテンス生成ステップと、

前記基本センテンスから、格フレームの生成に不要な語彙を削除する不要語彙削除ステップと、

前記不要語彙が削除された基本センテンスにおける動詞について、格フレームを生成する格フレーム生成ステップと、

同一の動詞についての格フレームに基づいて、その動詞の下位範疇化情報と項構造情報を生成し、前記補助情報として出力する補助情報生成ステップとを備えるプログラムが記録されていることを特徴とする記録媒体。

【請求項10】 入力文を自然言語処理する自然言語処理装置であって、

少なくとも、動詞の下位範疇化情報と項構造情報からなる補助情報を記憶している補助情報記憶手段と、

前記補助情報記憶手段から、前記入力文に含まれる動詞についての前記補助情報を検索する検索手段と、

前記入力文中に照応形が存在するかどうかを判定する判

10

20

30

40

50

定手段と、

前記入力文中に存在する照応形の属性を、その入力文に含まれる動詞についての前記補助情報に基づいて認識する属性認識手段と、

前記照応形の属性に基づいて、前記照応形が指し示す先行詞を決定する先行詞決定手段と、

前記先行詞決定手段において決定された先行詞を用いて、前記入力文の構文解析または意味解析を行う解析手段とを備えることを特徴とする自然言語処理装置。

【請求項 11】 前記判定手段は、前記入力文の構文解析結果、または前記入力文に含まれる動詞についての前記補助情報の下位範疇化情報に基づいて、前記入力文中に照応形が存在するかどうかを判定することを特徴とする請求項 10 に記載の自然言語処理装置。

【請求項 12】 前記照応形は、代名詞またはゼロ照応形であることを特徴とする請求項 10 に記載の自然言語処理装置。

【請求項 13】 対話履歴を記憶しながら、対話を行う対話装置であり、

前記先行詞決定手段は、前記対話履歴を参照することにより、前記先行詞を決定することを特徴とする請求項 10 に記載の自然言語処理装置。

【請求項 14】 ユーザに対して、前記先行詞の内容の問い合わせを行う問い合わせ手段をさらに備え、前記先行詞決定手段は、前記問い合わせに対するユーザの回答に基づいて、前記先行詞を決定することを特徴とする請求項 10 に記載の自然言語処理装置。

【請求項 15】 入力文を自然言語処理する自然言語処理方法であって、

少なくとも、動詞の下位範疇化情報と項構造情報からなる補助情報を記憶している補助情報記憶手段から、前記入力文に含まれる動詞についての前記補助情報を検索する検索ステップと、

前記入力文中に照応形が存在するかどうかを判定する判定ステップと、

前記入力文中に存在する照応形の属性を、その入力文に含まれる動詞についての前記補助情報に基づいて認識する属性認識ステップと、

前記照応形の属性に基づいて、前記照応形が指し示す先行詞を決定する先行詞決定ステップと、

前記先行詞決定ステップにおいて決定された先行詞を用いて、前記入力文の構文解析または意味解析を行う解析ステップとを備えることを特徴とする自然言語処理方法。

【請求項 16】 入力文を自然言語処理する自然言語処理を、コンピュータに行わせるプログラムであって、少なくとも、動詞の下位範疇化情報と項構造情報からなる補助情報を記憶している補助情報記憶手段から、前記入力文に含まれる動詞についての前記補助情報を検索する検索ステップと、

前記入力文中に照応形が存在するかどうかを判定する判定ステップと、

前記入力文中に存在する照応形の属性を、その入力文に含まれる動詞についての前記補助情報に基づいて認識する属性認識ステップと、

前記照応形の属性に基づいて、前記照応形が指し示す先行詞を決定する先行詞決定ステップと、

前記先行詞決定ステップにおいて決定された先行詞を用いて、前記入力文の構文解析または意味解析を行う解析ステップとを備えることを特徴とするプログラム。

【請求項 17】 入力文を自然言語処理する自然言語処理を、コンピュータに行わせるプログラムが記録されている記録媒体であって、

少なくとも、動詞の下位範疇化情報と項構造情報からなる補助情報を記憶している補助情報記憶手段から、前記入力文に含まれる動詞についての前記補助情報を検索する検索ステップと、

前記入力文中に照応形が存在するかどうかを判定する判定ステップと、

前記入力文中に存在する照応形の属性を、その入力文に含まれる動詞についての前記補助情報に基づいて認識する属性認識ステップと、

前記照応形の属性に基づいて、前記照応形が指し示す先行詞を決定する先行詞決定ステップと、

前記先行詞決定ステップにおいて決定された先行詞を用いて、前記入力文の構文解析または意味解析を行う解析ステップとを備えるプログラムが記録されていることを特徴とする記録媒体。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、自然言語処理装置および自然言語処理方法、並びにプログラムおよび記録媒体に関し、動詞について、その下位範疇化情報および項構造情報を得ることができるようし、さらに、その下位範疇化情報および項構造情報を用いて、照応形の先行詞を決定して、精度の高い対話や翻訳等の自然言語処理を行うことができるようにする自然言語処理装置および自然言語処理方法、並びにプログラムおよび記録媒体に関する。

【0002】

【従来の技術】従来の自然言語処理装置では、入力された文（入力文）が形態素解析され、さらに、その形態素解析結果に基づき、構文解析、意味解析が行われ、入力文の意味内容が理解される。そして、自然言語処理装置が、例えば、ユーザとの対話を行う対話装置である場合には、入力文の意味内容の理解に基づいて、その入力文に対する応答文が生成されて出力される。

【0003】

【発明が解決しようとする課題】ところで、例えば、

「もう食べましたか？」という入力文においては、食べ

たのは誰かという主語と、食べたものが何かという直接目的語が欠けている。従って、この入力文「もう食べましたか？」については、その欠けている主語と直接目的語を決定することができないと、その意味を正確に理解したということができない。

【0004】ここで、例えば、岩波講座-言語の科学6「生成文法」岩波書店、1997年や、橋田浩一「Global Document Annotation;GDA」電総研、1998年等の記載の自然言語理論によれば、ゼロ照応形(zero anaphora)と呼ばれる、表現されないが、目的語の位置にあり、照応関係を成立させる代名詞のようなものが存在する。即ち、この自然言語理論では、ある位置にあるべき名詞句が欠けている場合に、その位置に、ゼロ照応形(zero anaphora)が存在するとして扱われる。

【0005】なお、照応(anaphora)とは、代名詞、指示詞などの代用表現(照応形)とその指し示す対象(先行詞)との組によって表わされる言語現象であり、表現されない照応形が、ゼロ照応形である。

【0006】上述の入力文「もう食べましたか？」を正確に理解するためには、例えば、いま、ゼロ照応形を、proと表すこととすると、構文解析において、入力文「もう食べましたか？」における動詞「食べる」を、どのような構成素を必要とするものであるかを基準に分類し、その分類結果に基づき、入力文「もう食べましたか？」が、「pro(主語)もうpro(直接目的語)食べましたか？」であると分析(解析)する必要がある。さらには、ゼロ照応形(pro)が存在する場合には、そのゼロ照応形が指し示す先行詞が、具体的に何であるかを決定する必要がある。具体的には、入力文「もう食べましたか？」については、食べたのが誰であるのかと、食べたのか何であるのかを決定する必要がある。

【0007】ここで、動詞の種別としては、動作主(Agent)を主語にとる自動詞(intransitive)、対象(Theme)を主語にとる能格動詞(ergative)、直接目的語を選択する他動詞(transitive)、および直接目的語と間接目的語の両方を選択する二重目的語他動詞(ditransitive)の4つがあり、動詞を分類するとは、動詞を、これらの自動詞、能格動詞、他動詞、二重目的語他動詞のうちのいずれかに分類することを意味する。なお、上述の動詞「食べる」は他動詞である。

【0008】しかしながら、日本語においては、主語や目的語が頻繁に省略されるため、従来の自然言語処理装置では、構文解析時に、表層でも、また深層でも、動詞の分類、およびゼロ照応形を考慮した分析はあまり行われていなかった。

【0009】従って、従来の自然言語処理装置では、入力文におけるゼロ照応形の有無を判断することも、さらには、ゼロ照応形がある場合に、その先行詞を決定することもあまり行われていなかったため、精度の高い構文解析や意味解析を行うことができずに、入力文の意味を

正確に理解することができないことが多かった。

【0010】本発明は、このような状況に鑑みてなされたものであり、精度の高い構文解析や意味解析を可能とし、さらに、それにより、入力文の意味を正確に理解することができるようにするものである。

【0011】

【課題を解決するための手段】本発明の第1の自然言語処理装置は、コーパスデータの形態素解析結果から、格フレームの生成対象とする単位である基本センテンスを生成する基本センテンス生成手段と、基本センテンスから、格フレームの生成に不要な語彙を削除する不要語彙削除手段と、不要語彙が削除された基本センテンスにおける動詞について、格フレームを生成する格フレーム生成手段と、同一の動詞についての格フレームに基づいて、その動詞の下位範疇化情報と項構造情報を生成し、補助情報として出力する補助情報生成手段とを備えることを特徴とする。

【0012】本発明の第1の自然言語処理方法は、コーパスデータの形態素解析結果から、格フレームの生成対象とする単位である基本センテンスを生成する基本センテンス生成ステップと、基本センテンスから、格フレームの生成に不要な語彙を削除する不要語彙削除ステップと、不要語彙が削除された基本センテンスにおける動詞について、格フレームを生成する格フレーム生成ステップと、同一の動詞についての格フレームに基づいて、その動詞の下位範疇化情報と項構造情報を生成し、補助情報として出力する補助情報生成ステップとを備えることを特徴とする。

【0013】本発明の第1のプログラムは、コーパスデータの形態素解析結果から、格フレームの生成対象とする単位である基本センテンスを生成する基本センテンス生成ステップと、基本センテンスから、格フレームの生成に不要な語彙を削除する不要語彙削除ステップと、不要語彙が削除された基本センテンスにおける動詞について、格フレームを生成する格フレーム生成ステップと、同一の動詞についての格フレームに基づいて、その動詞の下位範疇化情報と項構造情報を生成し、補助情報として出力する補助情報生成ステップとを備えることを特徴とする。

【0014】本発明の第1の記録媒体は、コーパスデータの形態素解析結果から、格フレームの生成対象とする単位である基本センテンスを生成する基本センテンス生成ステップと、基本センテンスから、格フレームの生成に不要な語彙を削除する不要語彙削除ステップと、不要語彙が削除された基本センテンスにおける動詞について、格フレームを生成する格フレーム生成ステップと、同一の動詞についての格フレームに基づいて、その動詞の下位範疇化情報と項構造情報を生成し、補助情報として出力する補助情報生成ステップとを備えるプログラムが記録されていることを特徴とする。

【0015】本発明の第2の自然言語処理装置は、少なくとも、動詞の下位範疇化情報と項構造情報からなる補助情報を記憶している補助情報記憶手段から、入力文に含まれる動詞についての補助情報を検索する検索手段と、入力文中に照応形が存在するかどうかを判定する判定手段と、入力文中に存在する照応形の属性を、その入力文に含まれる動詞についての補助情報に基づいて認識する属性認識手段と、照応形の属性に基づいて、照応形が指し示す先行詞を決定する先行詞決定手段と、先行詞決定手段において決定された先行詞を用いて、入力文の構文解析または意味解析を行う解析手段とを備えることを特徴とする。

【0016】本発明の第2の自然言語処理方法は、少なくとも、動詞の下位範疇化情報と項構造情報からなる補助情報を記憶している補助情報記憶手段から、入力文に含まれる動詞についての補助情報を検索する検索ステップと、入力文中に照応形が存在するかどうかを判定する判定ステップと、入力文中に存在する照応形の属性を、その入力文に含まれる動詞についての補助情報に基づいて認識する属性認識ステップと、照応形の属性に基づいて、照応形が指し示す先行詞を決定する先行詞決定ステップと、先行詞決定ステップにおいて決定された先行詞を用いて、入力文の構文解析または意味解析を行う解析ステップとを備えることを特徴とする。

【0017】本発明の第2のプログラムは、少なくとも、動詞の下位範疇化情報と項構造情報からなる補助情報を記憶している補助情報記憶手段から、入力文に含まれる動詞についての補助情報を検索する検索ステップと、入力文中に照応形が存在するかどうかを判定する判定ステップと、入力文中に存在する照応形の属性を、その入力文に含まれる動詞についての補助情報に基づいて認識する属性認識ステップと、照応形の属性に基づいて、照応形が指し示す先行詞を決定する先行詞決定ステップと、先行詞決定ステップにおいて決定された先行詞を用いて、入力文の構文解析または意味解析を行う解析ステップとを備えることを特徴とする。

【0018】本発明の第2の記録媒体は、少なくとも、動詞の下位範疇化情報と項構造情報からなる補助情報を記憶している補助情報記憶手段から、入力文に含まれる動詞についての補助情報を検索する検索ステップと、入力文中に照応形が存在するかどうかを判定する判定ステップと、入力文中に存在する照応形の属性を、その入力文に含まれる動詞についての補助情報に基づいて認識する属性認識ステップと、照応形の属性に基づいて、照応形が指し示す先行詞を決定する先行詞決定ステップと、先行詞決定ステップにおいて決定された先行詞を用いて、入力文の構文解析または意味解析を行う解析ステップとを備えるプログラムが記録されていることを特徴とする。

【0019】本発明の第1の自然言語処理装置および自

然言語処理方法、並びにプログラムにおいては、コーパスデータの形態素解析結果から、格フレームの生成対象とする単位である基本センテンスが生成され、その基本センテンスから、格フレームの生成に不要な語彙が削除される。さらに、不要語彙が削除された基本センテンスにおける動詞について、格フレームが生成され、同一の動詞についての格フレームに基づいて、その動詞の下位範疇化情報と項構造情報が生成されて、補助情報として出力される。

10 【0020】本発明の第2の自然言語処理装置および自然言語処理方法、並びにプログラムにおいては、少なくとも、動詞の下位範疇化情報と項構造情報からなる補助情報を記憶している補助情報記憶手段から、入力文に含まれる動詞についての補助情報が検索される一方、入力文中に照応形が存在するかどうか判定され、入力文中に存在する照応形の属性が、その入力文に含まれる動詞についての補助情報に基づいて認識される。そして、照応形の属性に基づいて、照応形が指し示す先行詞が決定され、その先行詞を用いて、入力文の構文解析または意味解析が行われる。

20 【0021】

【発明の実施の形態】図1は、本発明を適用した自然言語処理装置の一実施の形態の構成例を示している。

【0022】この自然言語処理装置は、自然言語の構文解析や意味解析を補助する補助情報を、多量のコーパスデータから求める補助情報生成装置を構成している。

30 【0023】即ち、図1の補助情報生成装置としての自然言語処理装置は、多量のコーパスデータから、動詞についての格フレームを生成し、さらに、その格フレームから、動詞の下位範疇化情報(subcategorization)と項構造情報(argument structure)を含む補助情報を生成するようになっている。

【0024】ここで、例えば、平岡冠二・松本祐治(1994)「コーパスからの動詞の格フレーム獲得と名詞のクラスタリング」情報処理学会、自然言語処理研究会、NL-104や、春野雅彦(1995)「最小汎化とオッカムの原理を用いた動詞格フレーム学習」情報処理学会、自然言語処理研究会、NL-108、李航・安倍直樹(1996)「Learning Dependencies between Case Frame Slots」情報処理学会、自然言語処理研究会、NL-116には、同義関係情報を含むシソーラスと呼ばれる辞書を作成するための格フレームの自動生成方法が記載されているが、図1の補助情報生成装置において生成される格フレームは、下位範疇化情報と項構造情報を含む補助情報の作成を目的とする点で、シソーラスを作成する目的で格フレームを生成するのとは異なる。

【0025】また、補助情報を構成する下位範疇化情報は、例えば、HPSG(Head-Driven Phrase Structure Grammar - C.Pollard & I.Sag(1996) Head-Driven Phrase Structure Grammar. CSLI & University of Chicago

Press)や、J P S G (Japanese Phrase Structure Grammar - T. Gunji & K. Hasida (1998) Topics in Constraint-Based Grammar of Japanese. Kluwer Academic Publishers ; 郡司隆男「制約に基づく文法の連続量の概念を取り入れた拡張の研究」(平成12年)文部省研究成果報告書)等に記載されている汎用の自然言語処理理論において重要な役割を担うもので、次のような情報である。

【0026】即ち、動詞は、ある特定の構造や特定の統語的、意味的機能を有する構成素を要求するが、動詞を、その動詞が要求する構成素を基準に分類することは、下位範疇化(subcategorization)と呼ばれる。具体的には、例えば、動詞「食べる」は、「レストランで、うどんを、箸で食べました。」のように、名詞句(うどん+「を」)を構成素として必要とし、さらに、場所を表す名詞句(レストラン+「で」)や、手段を表す名詞句(箸+「で」)を、必要に応じて、構成素として伴う。このように、動詞が必要とする構成素を基準に、動詞を分類するのが、下位範疇化であり、下位範疇化によって動詞を分類する基準となる構成素に関する情報が、

下位範疇化情報である。

【0027】さらに、補助情報を構成する項構造情報とは、動詞が必然的に伴う、または必要に応じて伴う構成素が、どのような位置に現れ、どのような意味的な役割を担うのか等といった情報を意味する。

【0028】図1の補助情報生成装置は、コーパスデータベース1、前処理部2、格フレームデータベース3、格フレーム処理部4、および補助情報データベース5から構成されている。

【0029】コーパスデータベース1は、多量のコーパスデータを記憶している。なお、コーパスデータとしては、例えば、新聞記事等の文を採用することができる。

【0030】前処理部2は、形態素解析部11、基本センテンスパターン抽出部12、削除部13、格フレーム生成部14から構成され、補助情報を生成する前処理として、コーパスデータベース1に記憶された多量のコーパスデータから、格フレームを生成する処理を行う。

【0031】即ち、形態素解析部11は、コーパスデータベース1からコーパスデータを読み出し、形態素解析を行う。そして、形態素解析部11は、コーパスデータの形態素解析結果を、基本センテンスパターン抽出部12と格フレーム生成部14に供給する。なお、形態素解析部11による形態素解析結果は、必要に応じて、後述する格フレーム処理部4において参照することができるようになっている。

【0032】基本センテンスパターン抽出部12は、形態素解析部11から供給されるコーパスデータの形態素解析結果から、格フレームの生成対象とする単位である基本センテンスを生成(抽出)し、削除部13に供給する。即ち、基本センテンスパターン抽出部12は、原則

的には、形態素解析部11が出力する形態素解析結果のうち、句点の次の形態素から句点の直前の形態素までを、基本センテンスとして抽出し、削除部13に供給する。

【0033】削除部13は、基本センテンスパターン抽出部12から供給される基本センテンスから、格フレームの生成に不要な語彙を削除し、格フレーム生成部14に供給する。

【0034】格フレーム生成部14は、必要に応じて、形態素解析部11から供給されるコーパスデータの形態素解析結果を参照しながら、削除部13から供給される基本センテンスにおける動詞について、格フレームを生成し、格フレームデータベース3に供給する。

【0035】格フレームデータベース3は、前処理部2(を構成する格フレーム生成部14)から供給される格フレームを記憶するようになっている。

【0036】格フレーム処理部4は、格フレーム統合部21、動詞分類部22、下位範疇化情報生成部23、項構造情報生成部24、および補助情報生成部25から構成され、格フレームデータベース3から、同一の動詞についての格フレームを読み出し、その同一の動詞についての格フレーム等に基づいて、その動詞を分類するとともに、その下位範疇化情報と項構造情報を生成し、補助情報として出力する。

【0037】即ち、格フレーム統合部21は、格フレームデータベース3から、同一の動詞についての格フレームを読み出し、それらの格フレームを統合して、後述する統合格フレームとする。そして、格フレーム統合部21は、各動詞についての統合各フレームを、動詞分類部22、下位範疇化情報生成部23、および項構造情報生成部24に供給する。

【0038】動詞分類部22は、格フレーム統合部21から供給される統合格フレームに対応する動詞を、自動詞、能格動詞、他動詞、または二重目的語他動詞の4つの種別のうちのいずれかに分類し、その分類結果を表す分類情報を、下位範疇化情報生成部23と補助情報生成部25に供給する。

【0039】下位範疇化情報生成部23は、格フレーム統合部21から供給される統合格フレームと、動詞分類部22から供給される分類情報に基づいて、その統合格フレームに対応する動詞の下位範疇化情報を生成し、項構造情報生成部24と補助情報生成部25に供給する。

【0040】項構造情報生成部24は、格フレーム統合部21から供給される統合格フレームと、下位範疇化情報生成部23から供給される下位範疇化情報に基づいて、その統合格フレームに対応する動詞の項構造情報を生成し、補助情報生成部25に供給する。

【0041】補助情報生成部25は、各動詞について、動詞分類部22から供給される分類情報、下位範疇化情報生成部23から供給される下位範疇化情報、および項

構造情報生成部24から供給される項構造情報を対応付けて補助情報とし、補助情報データベース5に供給する。

【0042】補助情報データベース5は、補助情報生成部25から供給される各動詞についての補助情報を記憶するようにしている。

【0043】次に、図2は、形態素解析部11がコーパスデータを形態素解析することにより出力する形態素解析結果の例を示している。

【0044】なお、図2は、例えば、コーパスデータ「特に県内果実が数量で一八%増、金額で三四%増と伸びが目立った。」についての形態素解析結果を示している。

【0045】形態素解析結果は、形態素の見出し、読み(音韻)、シソーラス情報で構成され、シソーラス情報は、形態素の構文的な属性(フィーチャー)(構文属性)や、意味的な属性(意味属性)を含む。さらに、シソーラス情報は、形態素が動詞である場合には、その動詞の原形も含む。

【0046】ここで、図2において、1番目の形態素「特に」のシソーラス情報における属性[CAT Adv]のCATは、品詞を表す属性タグであり、従って、その後に続く情報が品詞であることを表す。CATの後に続くAdvは、品詞が副詞であることを表している。

【0047】また、形態素「特に」のシソーラス情報における属性[VAL 特に]のVALは、形態素の値(見出し)を表す属性タグであり、従って、その後に続く情報「特に」が、対応する形態素であることを表す。

【0048】2番目の形態素「県内果実」のシソーラス情報における属性[CAT Noun]は、品詞が名詞であることを表す。また、形態素「県内果実」のシソーラス情報における属性[cl Compound=CN+CN]のclは、クラスを表す属性タグであり、従って、その後に続く情報がクラスであることを表す。clの後に続くCompound=CN+CNは、クラスが、一般名詞(CN)と一般名詞(CN)とが結合した複合名詞であることを表す。さらに、形態素「県内果実」のシソーラス情報における属性[Sem food]のSemは、意味を表す属性タグであり、従って、その後に続く情報が意味であることを表す。Semの後に続くfoodは、形態素が食べ物を意味するものであることを表す。形態素「県内果実」のシソーラス情報における属性[VAL 県内果実]は、そのシソーラス情報が、形態素「県内果実」に対応するものであることを表す。

【0049】3番目の形態素「が」のシソーラス情報における属性[CAT Case]は、品詞が助詞(Case)であることを表し、属性[cl abstract]は、クラスが格助詞(abstract)であることを表す。さらに、属性[fx nominative]のfxは、形態素のファンクション(文法的役割)を表す属性タグであり、従って、属性[fx nominative]は、ファンクションが主格(nominative)であることを表す。属

性[VAL が]は、そのシソーラス情報が、形態素「が」に対応するものであることを表す。

【0050】4番目の形態素「数量」のシソーラス情報における属性[CAT Noun]は、品詞が名詞であることを表し、属性[cl CNoun]は、クラスが一般名詞(CNoun)であることを表す。属性[Sem amount]は、形態素「数量」が量(amount)を意味するものであることを表し、属性[VAL 数量]は、そのシソーラス情報が、形態素「数量」に対応するものであることを表す。

【0051】5番目の形態素「で」のシソーラス情報における属性[CAT Case]は、品詞が助詞であることを表し、属性[cl lexical]は、クラスが非格助詞(lexical)であることを表す。属性[fx instrument]は、ファンクションが道具(instrument)であることを表し、属性[VAL で]は、シソーラス情報が、形態素「で」に対応するものであることを表す。

【0052】6番目の形態素「一八%増」のシソーラス情報における属性[CAT Noun]は、品詞が名詞であることを表し、属性[cl Compound=Num+Classifier+suf]は、クラスが、数詞(Num)と助数詞(Classifier)と接尾語(suf)とからなる複合(名詞)であることを表す。属性[Sem increase]は、形態素「一八%増」が増加(increase)を意味するものであることを表し、属性[VAL 一八%増]は、シソーラス情報が、形態素「一八%増」に対応するものであることを表す。

【0053】7番目の形態素「、」のシソーラス情報における属性[CAT Punctuation]は、形態素「、」(の品詞)が記号(Punctuation)であることを表し、属性[cl comma]は、クラスがコンマ(comma)(読点)であることを表す。属性[VAL 、]は、シソーラス情報が、形態素「、」に対応するものであることを表す。

【0054】8番目の形態素「金額」のシソーラス情報における属性[CAT Noun]は、品詞が名詞であることを表し、属性[cl CNoun]は、クラスが一般名詞であることを表す。属性[Sem money]は、形態素「金額」がお金(money)を意味するものであることを表し、属性[VAL 金額]は、シソーラス情報が、形態素「金額」に対応するものであることを表す。

【0055】9番目の形態素「で」のシソーラス情報は、5番目の形態素「で」のものと同一である。

【0056】10番目の形態素「三四%増」のシソーラス情報は、属性[VAL 三四%増]を除き、6番目の形態素「一八%増」のシソーラス情報と同一である。

【0057】11番目の形態素「と」のシソーラス情報における属性[CAT Complementizer]は、品詞が補文をとる助詞(Complementizer)であることを表し、属性[cl proposition]は、クラスが文の引用(proposition)であることを表す。属性[VAL と]は、シソーラス情報が、形態素「と」に対応するものであることを表す。

【0058】12番目の形態素「伸び」のシソーラス情

報における属性[CAT Noun]は、品詞が名詞であることを表し、属性[cl CNoun]は、クラスが一般名詞であることを表す。属性[Sem increase]は、形態素「伸び」が増加を意味することを表し、属性[VAL 伸び]は、シソーラス情報が、形態素「伸び」に対応するものであることを表す。

【0059】13番目の形態素「が」のシソーラス情報は、3番目の形態素「が」のものと同一である。

【0060】14番目の形態素「目立った」のシソーラス情報における属性[CAT Verb]は、品詞が動詞(Verb)であることを表し、属性[cl active]は、クラスが能動(active)であることを表す。属性[fm finite]のfmは、フォームを表す属性タグであり、属性[fm finite]は、フォームが時制を伴う形(finite)であることを表す。属性[Conj (cl 2) (Stem 目立つ) (fm aff-past) (Polarity aff) (Ts past)]のConjは、活用を表す属性タグであり、属性(cl 2)は、活用がクラス2(cl 2)の活用であることを表す。ここで、形態素解析部11においては、動詞の活用が幾つかのクラスにクラス分けされており、クラス2の活用は、動詞の原形が子音で終わるということを表す。属性(Stem 目立つ)は、形態素「目立った」の原形(Stem)が「目立つ」であることを表す。なお、Stemは、動詞の原形を表す属性タグである。属性(fm aff-past)は、形態素「目立った」のフォーム(fm)が、肯定(aff (affirmation))で、かつ過去(past)であることを表し、属性(Polarity aff)は、形態素「目立った」の極性(Polarity)が肯定(aff)であることを表す。属性(Tspast)は、形態素「目立った」の時制(Ts)が過去(past)であることを表す。属性(Style (cl plain) (fm zero))のStyleは、スタイル(文体)を表す属性タグであり、属性(cl plain)は、スタイルのクラス(cl)が非丁寧形であること(いわゆる「ですます調」でないこと)を表す。属性(fm zero)は、スタイルのフォーム(fm)が原形のみ(zero)であることを表し、属性[VAL 目立った]は、シソーラス情報が、形態素「目立った」に対応するものであることを表す。

【0061】15番目の形態素「。」のシソーラス情報における属性[CAT Punctuation]は、形態素「。」(の品詞)が記号(Punctuation)であることを表し、属性[cl period]は、クラスがピリオド(period)(句点)であることを表す。属性[VAL 。]は、シソーラス情報が、形態素「。」に対応するものであることを表す。

【0062】次に、図3は、削除部13が、基本センテンスパターン抽出部12から供給される基本センテンスから、格フレームの生成に不要な語彙(以下、適宜、不要語彙という)として削除する語彙の例を示している。

【0063】削除部13は、基本センテンスから、次のような8種類の語彙を、不要語彙として削除する。

【0064】即ち、削除部13は、第1に、基本センテンスから、副詞を、不要語彙として削除する。副詞は、

図3(A)に示すように、形態素解析結果から、シソーラス情報が、{[CAT Adverb]}となっている形態素を検索することによって検出することができる。

【0065】削除部13は、第2に、基本センテンスから、例えば、「夏場の」などといった名詞+助詞「の」を、不要語彙として削除する。名詞+助詞「の」は、図3(B)に示すように、形態素解析結果から、シソーラス情報が、{[CAT Noun]・・・・}となっている形態素と、{[CAT Case][cl abstract][fx genitive][VAL の]}となっている形態素が連続している部分を検索することによって検出することができる。

【0066】なお、図3において(後述する図5においても同様)、括弧{}内の・・・は、他の属性が記述され得ることを意味する。

【0067】削除部13は、第3に、基本センテンスから、例えば、「日本での」などといった名詞+助詞+助詞「の」を、不要語彙として削除する。名詞+助詞+助詞「の」は、図3(C)に示すように、形態素解析結果から、シソーラス情報が、{[CAT Noun]・・・・}となっている形態素、{[CAT Case]・・・・}となっている形態素、および{[CAT Case][cl abstract][fx genitive][VAL の]}となっている形態素が連続している部分を検索することによって検出することができる。

【0068】削除部13は、第4に、基本センテンスから、形容詞を、不要語彙として削除する。形容詞は、図3(D)に示すように、形態素解析結果から、シソーラス情報が、{[CAT Adjective][cl stative]・・・・}となっている形態素を検索することによって検出することができる。なお、属性[CAT Adjective]は、品詞が形容詞(Adjective)であることを表し、属性[cl stative]は、クラスが状態(stative)であることを表す。

【0069】削除部13は、第5に、基本センテンスから、例えば、「決定的な」などといった名詞(形容動詞語幹)+「な」を、不要語彙として削除する。名詞(形容動詞語幹)+「な」は、図3(E)に示すように、形態素解析結果から、シソーラス情報が、{[CAT Noun]・・・・}となっている形態素と、{[CAT Verb][cl copula]・・・・[VAL な]}となっている形態素が連続している部分を検索することによって検出することができる。なお、属性[cl copula]は、クラスが連結詞であることを表す。

【0070】削除部13は、第6に、基本センテンスから、例えば、「工場に対する」などといった名詞+後置詞を、不要語彙として削除する。名詞+後置詞は、例えば、図3(F)に示すように、形態素解析結果から、シソーラス情報が、{[CAT Noun]・・・・}となっている形態素と、{[CAT Postposition]・・・・}となっている形態素が連続している部分を検索することによって検出することができる。なお、属性[CAT Postposition]は、品詞が後置詞(Postposition)であることを表す。

【0071】削除部13は、第7に、基本センテンスから、括弧で囲まれた部分を、不要語彙として削除する。括弧で囲まれた部分は、図3(G)に示すように、形態素解析結果から、シソーラス情報が、{[CAT Punctuation][cl L-]}となっている形態素から、{[CAT Punctuation][cl R-]}となっている形態素までの部分を検索することによって検出することができる。なお、属性[cl L-]は、クラスが括弧(例えば、"(")など)であることを表し、属性[cl R-]は、クラスが閉じ括弧(例えば、")")などであることを表す。

【0072】削除部13は、第8に、基本センテンスから、括弧で囲まれた部分+助詞「の」を、不要語彙として削除する。括弧で囲まれた部分+助詞「の」は、図3(H)に示すように、形態素解析結果から、シソーラス情報が、{[CAT Punctuation][cl L-]}となっている形態素から、{[CAT Punctuation][cl R-]}となっている形態素までの部分と、その後に、シソーラス情報が、{[CAT Case][cl abstract][fx genitive][VAL の]}となっている形態素を検索することによって検出することができる。

【0073】削除部13では、以上のような8種類の語彙が不要語彙として、基本センテンスから削除される。

【0074】従って、例えば、上述したコーパスデータ「特に県内果実が数量で一八%増、金額で三四%増と伸びが目立った。」については、削除部13からは、次のような基本センテンスが出力される。

【0075】即ち、コーパスデータ「特に県内果実が数量で一八%増、金額で三四%増と伸びが目立った。」については、基本センテンスパターン抽出部12において、そのコーパスデータから句点を除いた「特に県内果実が数量で一八%増、金額で三四%増と伸びが目立った」が、基本センテンスとして抽出される。そして、削除部13においては、「特に県内果実が数量で一八%増、金額で三四%増と伸びが目立った」から、図3

(A)の、品詞が副詞であることに該当する形態素「特に」が削除され、「特に県内果実が数量で一八%増、金額で三四%増と伸びが目立った」が出力される。

【0076】従って、図2に示したコーパスデータ「特に県内果実が数量で一八%増、金額で三四%増と伸びが目立った。」の形態素解析結果については、削除部13においては、図4に示すように、副詞である形態素「特に」と、句点である形態素「。」に関する情報がないものとなって出力される。

【0077】次に、格フレーム生成部14は、削除部13が出力する基本センテンスにおける動詞について、格フレームを生成するが、この格フレームの生成は、基本センテンスに含まれる動詞の「基準形」を、格フレームの見出しとして用いて行われるようになっている。即ち、格フレームは、その格フレームが、どのような動詞についてのものであるかを表す、その動詞の見出しと、

基本センテンスにおいて、その動詞が伴う助詞に関する情報とからなり、格フレームの見出しとしては、動詞の基準形が用いられる。

【0078】ここで、格フレームの見出しとなる動詞の基準形とは、例えば、図5に示すように定義されるものである。

【0079】即ち、以下説明する3つの例外を除いて、原則的には、基本センテンスに含まれる動詞の原形が、その動詞の基準形となる。具体的には、例えば、図5

10 (A)に示すように、基本センテンスに、動詞である形態素「目立つ」や「目立った」が含まれる場合には、その原形「目立つ」が基準形となる。

【0080】なお、動詞の原形は、図2で説明したように、形態素解析結果のシソーラス情報の中のStem属性タグとともに記述されているから、シソーラス情報を参照することで認識することができる。

【0081】次に、第1の例外として、基本センテンスに、サ変名詞+動詞「する」が含まれている場合には、動詞「する」の原形ではなく、サ変名詞+動詞「する」が、動詞の基準形となる。

【0082】従って、例えば、図5(B)に示すように、形態素解析結果のシソーラス情報が、{[CAT Noun][cl Vnoun]・・・[VAL 適用]}となっている形態素「適用」と、{[CAT Verb][cl active][fm finite]・・・(Stem する)(fm aff-non-past)・・・[VAL する]}となっている形態素「する」が連続する場合には、「適用する」が動詞の基準形とされる。なお、属性[cl Vnoun]は、クラスがサ変名詞(Vnoun)であることを表し、属性(fm aff-non-past)は、形態素「する」のフォーム(fm)が、肯定(aff)で、かつ過去でない(non-past)であることを表す。

【0083】第2の例外として、基本センテンスにおいて、動詞が2つ連続し、そのうちの最初の動詞が、シソーラス情報の中に、[fm infinite]と(pres.participle)の2つの属性を有する場合には、連続する2つの動詞のうちの最初の動詞の原形が、動詞の基準形となる。なお、属性[fm infinite]は、フォームが時制を伴わない形(infinite)であることを表し、属性(pres.participle)は、現在分詞(presentparticiple)であることを表す。

【0084】従って、例えば、図5(C)に示すように、基本センテンスにおいて、[fm infinite]と(pres.participle)の2つの属性を有する形態素「見込んで」に続いて、形態素「いる」があることにより、「見込んでいる」が存在する場合には、形態素「見込んで」の原形「見込む」が、動詞の基準形とされる。

【0085】第3の例外として、基本センテンスに、原形が「する」である動詞が含まれ、その動詞の直前に、サ変名詞がある場合は、サ変名詞+「する」が、動詞の基準形となる。

【0086】従って、例えば、図5(D)に示すように、形態素解析結果のシソーラス情報が、{[CAT Noun] [cl Vnoun]・・・[VAL 展開]}となっている形態素「展開」、{[CAT Verb]・・・[fm infinite]・・・(Stem する)(fm pres.participle)・・・[VAL して]}となっている形態素「して」、および{[CAT Verb]・・・[fm finite]・・・(Stem いる)・・・[VAL いる]}となっている形態素「いる」が連続している場合には、サ変名詞「展開」+「する」、即ち、「展開する」が、動詞の基準形とされる。

【0087】次に、図6は、格フレーム生成部14が作成する格フレームを示している。

【0088】図6は、動詞「目立つ」について、4つの基本センテンスからそれぞれ生成された4つの格フレーム{目立つ C_FRAME:で[instrument], が[increase]}, {目立つ C_FRAME:が[thing]}, {目立つ C_FRAME:と[proposition], が[thing]}, {目立つ C_FRAME:で[instrument], に[locative], が[increase]}を示している。

【0089】格フレームの先頭の文字列は、その格フレームに対応する動詞の見出しを表しており、この動詞の見出しとしては、図5で説明した動詞の基準形が用いられる。

【0090】また、格フレームにおけるC_FRAMEは、助詞(格助詞)を表すタグで、その後には、その見出しになっている動詞が、基本センテンスにおいてとっている助詞が記述される。なお、格フレームには、1以上の助詞を記述することができる。

【0091】さらに、格フレームにおける助詞の直後には、括弧[]が記述されるが、この括弧[]内には、その助詞のファンクション、またはその助詞の直前の形態素の意味が、その助詞の属性として記述される。なお、助詞のファンクションは、形態素解析結果におけるシソーラス情報のfx属性タグを検索することにより認識することができ、また、助詞の直前の形態素の意味は、シソーラス情報のSem属性タグを検索することにより認識することができる。

【0092】ここで、図6における1行目の格フレーム{目立つ C_FRAME:で[instrument], が[increase]}が、上述のコーパスデータ「特に県内果実が数量で一八%増、金額で三四%増と伸びが目立った。」について、格フレーム生成部14が後述する図12の格フレーム生成処理を行うことにより生成されるものである。

【0093】次に、図7は、格フレーム統合部21が、同一の動詞についての格フレームを統合することにより生成する統合格フレームを示している。

【0094】例えば、動詞(の基準形)「目立つ」について、図6に示したような4つの格フレームが得られている場合には、その4つの格フレームが統合されることにより、動詞「目立つ」について、図7に示したような

統合格フレームが生成される。

【0095】即ち、この場合、格フレーム統合部21は、動詞「目立つ」についての4つの格フレームに対する動詞の見出し「目立つ」を、統合格フレームの見出しとして配置し、続けて、その動詞の読みを配置する。なお、動詞の読みは、格フレーム統合部21が形態素解析部11の形態素解析結果を参照することで認識される。

【0096】さらに、格フレーム統合部21は、4つの格フレームの助詞と属性の、いわば論理和をとったものを求めて、タグsubcatとともに、統合格フレームに配置する。

【0097】即ち、図6に示した4つの格フレームには、「で」、「が」、「と」、「に」の4種類の助詞が存在するから、格フレーム統合部21は、この4種類の助詞「で」、「が」、「と」、「に」を、タグsubcatの後に配置する。さらに、図6の4つの格フレームにおいて、助詞「で」については、属性[instrument]しか存在しないので、統合格フレームにおける助詞「で」の後には、その属性[instrument]だけが配置される。また、図6の4つの格フレームにおいて、助詞(格助詞)「が」については、属性[increase]と[thing]の2種類が存在するので、統合格フレームにおける助詞「が」の後には、その2つの属性[increase]と[thing]が配置される。さらに、図6の4つの格フレームにおいて、助詞「と」については、属性[proposition]しか存在しないので、統合格フレームにおける助詞「と」の後には、その属性[proposition]だけが配置される。また、図6の4つの格フレームにおいて、助詞「に」については、属性[locative]しか存在しないので、統合格フレームにおける助詞「に」の後には、その属性[locative]だけが配置される。

【0098】次に、図8は、補助情報生成部25が、各動詞について生成する補助情報を示している。

【0099】図8は、動詞「目立つ」についての補助情報を示しており、その先頭と2番目には、図7に示した統合格フレームと同様に、動詞「目立つ」の見出し(動詞の基準形)と読みが配置される。

【0100】補助情報において、動詞の読みの後には、その動詞が、自動詞、能格動詞、他動詞、または二重目的語他動詞のうちのいずれに分類されるものであるかを表す分類情報が配置される。図8において、動詞「目立つ」は、対象(Theme)を主語にとる能格動詞であり、従って、分類情報としては、「能格動詞」が配置されている。なお、分類情報は、動詞分類部22から補助情報生成部25に供給されるものである。

【0101】補助情報において、分類情報の後には、動詞の下位範疇化情報が配置される。下位範疇化情報は、図8に示したように、下位範疇化情報であることを表すタグSUBCATとともに、例えば、<SUBCAT:NP[nom]>といった形で記述される。なお、NPは、名詞句を表し、[nom]

は、主格を表す。そして、下位範疇化情報<SUBCAT:NP[n om]>は、主格となる名詞句を必然的に伴うことを表す。この下位範疇化情報は、下位範疇化情報生成部23から補助情報生成部25に供給されるものである。

【0102】下位範疇化情報の後には、動詞の項構造情報が配置される。項構造情報は、図8に示したように、項構造情報であることを表すタグArgStrとともに、例えば、<ArgStr:Theme(thing/increase)-(Instrument)-(Locative)-(Proposition)>といった形で記述される。項構造情報(のArgStr:以降の記述)のうち、小括弧()や、中括弧{}で囲まれていない部分(以下、適宜、主情報という)は、下位範疇化情報において、動詞が必然的に伴うとされている構成素を表す。図8では、対象物を表すThemeが、主情報となっており、従って、下位範疇化情報も考慮すれば、図8の補助情報は、動詞「目立つ」が必然的に伴う、主格となる名詞句は、対象物であることを表す。

【0103】主情報の後の、中括弧{}内の記述は、その主情報の属性(シソーラス)を表す。図8における{thing/increase}のthingとincreaseは、それぞれ、物と増加を表し、従って、属性{thing/increase}は、主情報「Theme」が表す対象物が、物または増加を表すものであることを表す。

【0104】項構造情報の小括弧()内の記述は、動詞が必要に応じて伴うことのできる表現(語彙)の属性を表す。図8においては、道具を表すInstrument、場所を表すLocation、および文(埋め込み文)を表すPropositionが記述されており、従って、図8の補助情報は、動詞「目立つ」が、道具を表す表現、場所を表す表現、文を指し示す表現を、必要に応じて伴うことを表す。

【0105】次に、図9のフローチャートを参照して、図1の補助情報生成装置が行う自然言語処理としての、補助情報を生成する補助情報生成処理について説明する。

【0106】まず最初に、ステップS1において、形態素解析部11は、コーパスデータベース1に記憶されている多量のコーパスデータを順次読み出し、各コーパスデータについて、形態素解析を行う。形態素解析部11が、各コーパスデータについて形態素解析を行うことにより得られる形態素解析結果は、基本センテンスパターン抽出部12および格フレーム生成部14、並びに格フレーム処理部4に供給される。

【0107】その後、ステップS2に進み、基本センテンスパターン抽出部12は、形態素解析部11から供給される、各コーパスデータのついでに形態素解析結果から、基本センテンスを抽出する基本センテンスパターン抽出処理を行い、その結果得られる基本センテンスを、削除部13に供給して、ステップS3に進む。ステップS3では、削除部13が、基本センテンスパターン抽出部12から供給される各基本センテンスから不要語彙を

削除する不要語彙削除処理を行い、その不要語彙を削除した基本センテンスを、格フレーム生成部14に供給して、ステップS4に進む。ステップS4では、格フレーム生成部14は、削除部13から供給される各基本センテンスに関し、その基本センテンスに含まれる動詞について、格フレームを生成する格フレーム生成処理を行う。さらに、格フレーム生成部14は、その格フレーム生成処理によって生成した格フレームを、格フレームデータベース3に供給して記憶させ、ステップS5に進む。

【0108】ステップS5では、格フレーム統合部21が、格フレームデータベース3に記憶された格フレームから、同一の動詞についてのものを収集し、図6および図7で説明したように、その同一の動詞についての1以上の格フレームを統合して、統合格フレームを生成する。そして、格フレーム統合部21は、統合格フレームを、動詞分類部22、下位範疇化情報生成部23、項構造情報生成部24に供給して、ステップS6に進む。

【0109】ステップS6では、動詞分類部22が、格フレーム統合部21から供給される統合格フレームに基づいて、各統合格フレームに対応する動詞を、自動詞、能格動詞、他動詞、二重目的語他動詞のいずれかに分類し、その分類結果を表す分類情報を出力する動詞分類処理を行う。さらに、ステップS6では、下位範疇化情報生成部23が、格フレーム統合部21から供給される統合格フレーム、および動詞分類部22から供給される分類情報に基づいて、各統合格フレームに対応する動詞の下位範疇化情報を生成して出力する下位範疇化情報生成処理を行う。また、ステップS6では、項構造情報生成部24が、格フレーム統合部21から供給される統合格フレーム、および下位範疇化情報生成部23から供給される下位範疇化情報に基づいて、各統合格フレームに対応する動詞の項構造情報を生成して出力する項構造情報生成処理を行う。

【0110】その後、ステップS7に進み、補助情報生成部25が、動詞分類部22から供給される分類情報、下位範疇化情報生成部23から供給される下位範疇化情報、および項構造情報生成部24から供給される項構造情報を用い、各統合格フレームに対応する動詞について、図8に示したような補助情報を生成する。さらに、補助情報生成部25は、補助情報を、補助情報データベース5に供給して記憶させ、補助情報生成処理を終了する。

【0111】次に、図10のフローチャートを参照して、図1の基本センテンスパターン抽出部12が図9のステップS2で行う基本センテンスパターン抽出処理について説明する。

【0112】基本センテンスパターン抽出部12は、ステップS11において、その内蔵するバッファ(図示せず)をクリアするとともに、形態素解析部11において

10

20

30

40

50

形態素解析結果が得られたコーパスデータのうち、まだ処理の対象としていない最も古いものを注目コーパスデータとする。そして、ステップS12に進み、基本センテンスパターン抽出部12は、注目コーパスデータの形態素の、まだ読み込んでいない、より文頭に近いものを、注目形態素として、その形態素解析結果を読み込み、ステップS13に進む。ステップS13では、基本センテンスパターン抽出部12は、注目形態素が、句点であるかどうかを、その形態素解析結果を参照することによって判定する。

【0113】ステップS13において、注目形態素が句点でないと判定された場合、ステップS14に進み、基本センテンスパターン抽出部12は、注目形態素の形態素解析結果を、その内蔵するバッファに追加記憶させ、ステップS12に戻り、いま注目形態素となっている次の形態素を、新たな注目形態素として、以下、同様の処理を繰り返す。

【0114】また、ステップS13において、注目形態素が句点であると判定された場合、ステップS15に進み、基本センテンスパターン抽出部12は、その内蔵するバッファを参照することにより、注目形態素である句点の直前の形態素（あるいは句点以前にある最初の動詞）が、時制を伴う動詞であるかどうかを判定する。ステップS15において、注目形態素である句点の直前の形態素が、時制を伴う動詞でないと判定された場合、ステップS16およびS17をスキップして、ステップS18に進む。

【0115】また、ステップS15において、注目形態素である句点の直前の形態素が、時制を伴う動詞であると判定された場合、ステップS16に進み、基本センテンスパターン抽出部12は、その内蔵するバッファに、注目形態素である句点の直前の形態素以外に、時制を伴う動詞（の形態素解析結果）が記憶されていないかどうかを判定する。

【0116】ステップS16において、基本センテンスパターン抽出部12の内蔵するバッファに、注目形態素である句点の直前の形態素以外に、時制を伴う動詞が記憶されていると判定された場合、ステップS17をスキップして、ステップS18に進む。

【0117】一方、ステップS16において、基本センテンスパターン抽出部12の内蔵するバッファに、注目形態素である句点の直前の形態素以外に、時制を伴う動詞が記憶されていないと判定された場合、ステップS17に進み、基本センテンスパターン抽出部12は、その内蔵するバッファに記憶された形態素（解析結果）のシーケンスを、基本センテンスとして抽出し（読み出し）、削除部13に供給して、ステップS18に進む。

【0118】ステップS18では、基本センテンスパターン抽出部12は、まだ、注目コーパスデータとしていないコーパスデータがあるかどうかを判定する。ステッ

プS18において、まだ、注目コーパスデータとしていないコーパスデータがあると判定された場合、ステップS11に戻り、まだ、注目コーパスデータとしていないコーパスデータの1つが、新たに、注目コーパスデータとされ、以下、同様の処理が繰り返される。

【0119】また、ステップS18において、まだ、注目コーパスデータとしていないコーパスデータがないと判定された場合、基本センテンスパターン抽出処理を終了する。

10 【0120】以上のような基本センテンスパターン抽出処理によれば、句点の直後の形態素から、次の句点の直前の形態素までの形態素列であって、時制を伴う動詞を1つしか含んでいないもの（基本的には、単文）が、基本センテンスとして抽出される。

【0121】次に、図11のフローチャートを参照して、図1の削除部13が図9のステップS3で行う不要語彙削除処理について説明する。

20 【0122】削除部13は、まず最初に、ステップS21において、基本センテンスパターン抽出部12から供給される基本センテンスのうち、まだ、注目基本センテンスとしていないもののうちの1つを、注目基本センテンスとして、その注目基本センテンスを構成する形態素の数を、変数Nにセットする。

【0123】そして、削除部13は、ステップS22に進み、基本センテンスの形態素をカウントする変数iとjを、いずれも1に初期化し、ステップS23に進む。

30 【0124】ステップS23では、削除部13は、注目基本センテンスの先頭からi番目の形態素から、j番目の形態素までの形態素列を、変数Stringにセットし、ステップS24に進む。

【0125】ステップS24では、削除部13は、変数Stringにセットされている形態素列（または形態素）が、削除条件に該当するかどうかを判定する。

【0126】ここで、削除条件に該当する場合とは、図3で説明した不要語彙のいずれかに該当することを意味する。

【0127】ステップS24において、変数Stringにセットされている形態素列が削除条件に該当しないと判定された場合、ステップS25をスキップして、ステップS26に進む。また、ステップS24において、変数Stringにセットされている形態素列が削除条件に該当すると判定された場合、ステップS25に進み、削除部13は、その内蔵するバッファ（図示せず）に、変数Stringにセットされている形態素列を、削除対象としてバッファリングして、ステップS26に進む。

40 【0128】ステップS26では、削除部13が、変数jが、注目基本センテンスを構成する形態素の数Nに等しいかどうかを判定する。ステップS26において、変数jがNに等しくないと判定された場合、ステップS27に進み、削除部13は、変数jを1だけインクリメン

トして、ステップS 2 3に戻り、以下、同様の処理を繰り返す。

【0 1 2 9】また、ステップS 2 6において、変数jがNに等しいと判定された場合、ステップS 2 8に進み、削除部1 3は、変数iがNに等しいかどうかを判定する。ステップS 2 8において、変数iがNに等しくない

と判定された場合、ステップS 2 9に進み、削除部1 3は、変数iを1だけインクリメントするとともに、変数jに、変数iにセットされている値をセットして、ステップS 2 3に戻り、以下、同様の処理を繰り返す。

【0 1 3 0】一方、ステップS 2 8において、変数iがNに等しいと判定された場合、即ち、基本センテンスを構成する任意の形態素と形態素列について、不要語彙かどうかの判定を行った場合、ステップS 3 0に進み、削除部1 3は、注目基本センテンスから、その内蔵するバッファに削除対象として記憶されている形態素と形態素列を削除し、格フレーム生成部1 4に供給して、ステップS 3 1に進む。

【0 1 3 1】ステップS 3 1では、削除部1 3は、まだ、注目基本センテンスとしていない基本センテンスがあるかどうかを判定する。ステップS 3 1において、まだ、注目基本センテンスとしていない基本センテンスがあると判定された場合、ステップS 2 1に戻り、削除部1 3は、まだ、注目基本センテンスとしていない基本センテンスのうちの1つを、新たな注目基本センテンスとし、以下、同様の処理を繰り返す。

【0 1 3 2】また、ステップS 3 1において、まだ、注目基本センテンスとしていない基本センテンスがないと判定された場合、不要語彙削除処理を終了する。

【0 1 3 3】次に、図1 2のフローチャートを参照して、図1の格フレーム生成部1 4が図9のステップS 5で行う格フレーム生成処理について説明する。

【0 1 3 4】格フレーム生成部1 4は、まず最初に、ステップS 4 1において、削除部1 3から供給される基本センテンスのうち、まだ、注目基本センテンスとしていないもののうちの1つを、注目基本センテンスとして、その注目基本センテンスに含まれる動詞（以下、適宜、注目動詞という）の基準形を、その注目動詞についての格フレームの見出しとして記述する。

【0 1 3 5】そして、格フレーム生成部1 4は、ステップS 4 2に進み、基本センテンスの形態素をカウントする変数iを1に初期化し、ステップS 4 3に進む。

【0 1 3 6】ステップS 4 3では、格フレーム生成部1 4は、注目基本センテンスの最後からi番目の形態素を、変数Stringにセットし、ステップS 4 4に進む。

【0 1 3 7】ステップS 4 4では、格フレーム生成部1 4は、変数Stringにセットされている形態素が助詞であるかどうかを、その形態素解析結果のシソーラス情報（図2）を参照することにより判定する。

【0 1 3 8】ステップS 4 4において、変数Stringにセ

ットされている形態素が助詞でないと判定された場合、ステップS 4 5およびS 4 6をスキップして、ステップS 4 7に進む。

【0 1 3 9】また、ステップS 4 4において、変数Stringにセットされている形態素が助詞であると判定された場合、ステップS 4 5に進み、格フレーム生成部1 4は、変数Stringにセットされている助詞と、その属性を、注目動詞についての格フレームに記述し、ステップS 4 6に進む。なお、格フレーム生成部1 4は、助詞の属性を、形態素解析部1 1による形態素解析結果のシソーラス情報を参照することで認識する。

【0 1 4 0】ステップS 4 6では、格フレーム生成部1 4が、変数Stringにセットされている助詞が、注目基本センテンスの最後から数えて、1つ目の「は」、または2つ目の「が」、「に」、若しくは「を」のうちのいずれかに該当するかどうかを判定する。

【0 1 4 1】ステップS 4 6において、変数Stringにセットされている助詞が、注目基本センテンスの最後から数えて、1つ目の「は」、2つ目の「が」、2つ目の「に」、または2つ目の「を」のうちのいずれかに該当すると判定された場合、ステップS 4 7をスキップして、ステップS 4 9に進む。

【0 1 4 2】また、ステップS 4 6において、変数Stringにセットされている助詞が、注目基本センテンスの最後から数えて、1つ目の「は」、2つ目の「が」、2つ目の「に」、および2つ目の「を」のうちのいずれにも該当しないと判定された場合、ステップS 4 7に進み、格フレーム生成部1 4は、変数Stringにセットされている形態素が、注目基本センテンスの先頭の形態素であるかどうかを判定する。

【0 1 4 3】ステップS 4 7において、変数Stringにセットされている形態素が、注目基本センテンスの先頭の形態素でないと判定された場合、ステップS 4 8に進み、格フレーム生成部1 4は、変数iを1だけインクリメントして、ステップS 4 3に戻り、以下、同様の処理を繰り返す。

【0 1 4 4】また、ステップS 4 7において、変数Stringにセットされている形態素が、注目基本センテンスの先頭の形態素であると判定された場合、ステップS 4 9に進み、格フレーム生成部1 4は、まだ、注目基本センテンスとしていない基本センテンスがあるかどうかを判定する。ステップS 4 9において、まだ、注目基本センテンスとしていない基本センテンスがあると判定された場合、ステップS 4 1に戻り、格フレーム生成部1 4は、まだ、注目基本センテンスとしていない基本センテンスのうちの1つを、新たな注目基本センテンスとし、以下、同様の処理を繰り返す。

【0 1 4 5】また、ステップS 4 9において、まだ、注目基本センテンスとしていない基本センテンスがないと判定された場合、格フレーム生成処理を終了する。

【0146】以上のような格フレーム生成処理によれば、削除部13が出力する基本センテンスの文末から文頭方向に辿っていき、1つ目の「は」、2つ目の「が」、2つ目の「に」、または2つ目の「を」のうちのいずれかに到達するまでに現れる助詞とその属性が、その基本センテンスに含まれる動詞についての格フレームに記述され、これにより、図6に示したような格フレームが生成される。

【0147】次に、図13のフローチャートを参照して、図1の動詞分類部22が図9のステップS6で行う動詞分類処理について説明する。

【0148】動詞分類部22は、ステップS61において、格フレーム統合部21が出力する統合格フレームのうち、まだ、注目統合格フレームとしていないものの1つを注目統合格フレームとし、その注目統合格フレームから、サブカテゴリ情報を読み出す。

【0149】ここで、サブカテゴリ情報とは、図7に示した統合格フレームにおいて、subcatタグ以降に記述される情報を意味する。

【0150】その後、ステップS62に進み、動詞分類部22は、注目統合格フレームが、そのサブカテゴリ情報に、格助詞「を」を含まないが、格助詞「が」を含み、かつ、その格助詞「が」と名詞とで構成される名詞+格助詞「が」が、注目統合格フレームに対応する動詞の動作主(agent)になり得るという自動詞が満たす条件(以下、適宜、自動詞条件という)を満たすかどうかを判定する。

【0151】ここで、名詞+格助詞「が」が、注目統合格フレームに対応する動詞の動作主になり得るかどうかは、その動詞を含むコーパスデータの形態素解析結果におけるシソーラス情報の意味を表すSemタグを参照することで判定することができる。

【0152】ステップS62において、注目統合格フレームが、自動詞条件を満たすと判定された場合、ステップS63に進み、動詞分類部22は、注目統合格フレームに対応する動詞(注目統合格フレームの見出しとなっている動詞)を、自動詞に分類し、その旨を表す分類情報を、下位範疇化情報生成部23と補助情報生成部25に供給して、ステップS71に進む。

【0153】また、ステップS62において、注目統合格フレームが、自動詞条件を満たさないと判定された場合、ステップS64に進み、動詞分類部22は、注目統合格フレームが、そのサブカテゴリ情報に、格助詞「を」を含まないが、格助詞「が」を含み、かつ、その格助詞「が」と名詞とで構成される名詞+格助詞「が」が、注目統合格フレームに対応する動詞の動作主(agent)になり得ないという能格動詞が満たす条件(以下、適宜、能格動詞条件という)を満たすかどうかを判定する。

【0154】ステップS64において、注目統合格フレーム

ームが、能格動詞条件を満たすと判定された場合、ステップS65に進み、動詞分類部22は、注目統合格フレームに対応する動詞を、能格動詞に分類し、その旨を表す分類情報を、下位範疇化情報生成部23と補助情報生成部25に供給して、ステップS71に進む。

【0155】また、ステップS64において、注目統合格フレームが、能格動詞条件を満たさないと判定された場合、ステップS66に進み、動詞分類部22は、注目統合格フレームが、そのサブカテゴリ情報に、格助詞「を」を含むが、間接目的語をとるのに必要な助詞「に」を含まないという他動詞が満たす条件(以下、適宜、他動詞条件という)を満たすかどうかを判定する。

【0156】ステップS66において、注目統合格フレームが、他動詞条件を満たすと判定された場合、ステップS67に進み、動詞分類部22は、注目統合格フレームに対応する動詞を、他動詞に分類し、その旨を表す分類情報を、下位範疇化情報生成部23と補助情報生成部25に供給して、ステップS71に進む。

【0157】また、ステップS66において、注目統合格フレームが、他動詞条件を満たさないと判定された場合、ステップS68に進み、動詞分類部22は、注目統合格フレームが、そのサブカテゴリ情報に、格助詞「を」を含み、さらに、間接目的語をとるのに必要な助詞「に」を含むという二重目的語他動詞が満たす条件(以下、適宜、二重目的語他動詞条件という)を満たすかどうかを判定する。

【0158】ステップS68において、注目統合格フレームが、二重目的語他動詞条件を満たすと判定された場合、ステップS69に進み、動詞分類部22は、注目統合格フレームに対応する動詞を、二重目的語他動詞に分類し、その旨を表す分類情報を、下位範疇化情報生成部23と補助情報生成部25に供給して、ステップS71に進む。

【0159】また、ステップS68において、注目統合格フレームが、二重目的語他動詞条件を満たさないと判定された場合、ステップS70に進み、例えば、注目統合格フレームを、格フレーム処理部4における処理対象から除外する等のエラー処理を行い、ステップS71に進む。

【0160】ステップS71では、動詞分類部22が、まだ、注目統合格フレームとしていない統合格フレームがあるかどうかを判定する。ステップS71において、まだ、注目統合格フレームとしていない統合格フレームがあると判定された場合、ステップS61に戻り、動詞分類部22は、まだ、注目統合格フレームとしていない統合格フレームのうちの1つを、新たな注目統合格フレームとし、以下、同様の処理を繰り返す。

【0161】また、ステップS71において、まだ、注目統合格フレームとしていない統合格フレームがないと判定された場合、動詞分類処理を終了する。

【0162】次に、図14のフローチャートを参照して、図1の下位範疇化情報生成部23が図9のステップS6で行う下位範疇化情報生成処理について説明する。

【0163】下位範疇化情報生成部23は、まず最初に、ステップS81において、格フレーム統合部21が出力する統合格フレームのうち、まだ、注目統合格フレームとしていないものの1つを注目統合格フレームとして受信し、さらに、その注目統合格フレームについて、動詞分類部22が出力する分類情報を受信する。

【0164】そして、ステップS82に進み、下位範疇化情報生成部23は、注目統合格フレームと、その分類情報に基づいて、注目統合格フレームに対応する動詞の下位範疇化情報を生成する。

【0165】即ち、下位範疇化情報生成部23は、注目統合格フレームに対応する動詞（以下、適宜、注目動詞という）の分類情報から、その注目動詞が、自動詞、能格動詞、他動詞、または二重目的語他動詞のうちのいずれであるかを認識し、その認識結果と、注目統合格フレームから、注目動詞が必然的に伴う構成素を認識する

（注目動詞が、上述の4つの動詞のうちのいずれであるかによって、その注目動詞が必然的に伴う構成素に制約をかけ、その制約の下で、注目統合格フレームから、注目動詞が必然的に伴う構成素を認識する）。そして、下位範疇化情報生成部23は、その注目動詞が必然的に伴う構成素に関する情報を、下位範疇化情報として、項構造情報生成部24と補助情報生成部25に出力する。

【0166】従って、例えば、いま、図7に示した動詞「目立つ」についての統合格フレームが注目統合格フレームとされたとした場合を考えると、まず、動詞「目立つ」は、上述したように、能格動詞であり、主格となる名詞句を必然的に伴う。また、図7に示した動詞「目立つ」についての統合格フレームにおいては、主格を表す格助詞「が」だけが存在し、他の格助詞は存在しない。そこで、下位範疇化情報生成部23では、主格となる名詞句を必然的に伴うことを表すNP[nom]が、動詞「目立つ」の下位範疇化情報として生成される。なお、図8で説明したように、NPは名詞句を表し、[nom]は主格を表す。

【0167】その後、ステップS83に進み、下位範疇化情報生成部23が、まだ、注目統合格フレームとしていない統合格フレームがあるかどうかを判定する。ステップS83において、まだ、注目統合格フレームとしていない統合格フレームがあると判定された場合、ステップS81に戻り、下位範疇化情報生成部23は、まだ、注目統合格フレームとしていない統合格フレームのうちの1つを、新たな注目統合格フレームとし、以下、同様の処理を繰り返す。

【0168】また、ステップS83において、まだ、注目統合格フレームとしていない統合格フレームがないと判定された場合、下位範疇化情報生成処理を終了する。

【0169】次に、図15のフローチャートを参照して、図1の項構造情報生成部24が図9のステップS6で行う項構造情報生成処理について説明する。

【0170】項構造情報生成部24は、まず最初に、ステップS91において、格フレーム統合部21が出力する統合格フレームのうち、まだ、注目統合格フレームとしていないものの1つを注目統合格フレームとして受信し、さらに、その注目統合格フレームについて、下位範疇化情報生成部23が出力する下位範疇化情報を受信する。

【0171】そして、ステップS92に進み、項構造情報生成部24は、注目統合格フレームと、その下位範疇化情報に基づいて、注目統合格フレームに対応する動詞が必然的に伴う（必須）の格助詞と、その属性を認識する。

【0172】即ち、項構造情報生成部24は、注目統合格フレームに対応する動詞（以下、適宜、注目動詞という）の下位範疇化情報から、その注目動詞に必須の格助詞を認識し、さらに、その格助詞の属性を、注目統合格フレームから認識する。

【0173】従って、例えば、いま、図7に示した動詞「目立つ」についての統合格フレームが注目統合格フレームとされたとした場合、下位範疇化情報としては、上述したように、主格となる名詞句を必然的に伴うことを表すNP[nom]が生成されるから、図7の注目統合格フレームに記述された助詞「で」、「が」、「に」、「と」のうち、主格を表す格助詞「が」が、注目動詞「目立つ」に必須の格助詞として認識される。さらに、図7の注目統合格フレームにおいては、格助詞「が」の属性として、その格助詞「が」とともに主格を構成する名詞が、動作主(agent)となり得ない属性[increase]または[thing]を有するものとなっているから、それらの上位概念としての、例えば、対象物を表す属性Themeが認識され、その属性Themeが、下位概念として、属性[increase]と[thing]を含むことを表す属性Theme(thing/increase)が、注目動詞「目立つ」に必須の格助詞の属性として認識される。

【0174】その後、ステップS93に進み、項構造情報生成部24は、注目統合格フレームと、その下位範疇化情報に基づいて、注目統合格フレームに対応する動詞が必要に応じて伴う助詞（以下、適宜、オプションの助詞という）と、その属性を認識する。

【0175】即ち、項構造情報生成部24は、注目統合格フレームに記述された助詞から、ステップS92で認識した必須の格助詞を除いたものを、オプションの助詞として認識する。さらに、項構造情報生成部24は、注目統合格フレームにおいて、オプションの助詞として認識した助詞に付されている属性を、オプションの助詞の属性として認識する。

【0176】従って、例えば、いま、図7に示した動詞

「目立つ」についての統合合格フレームが注目統合合格フレームとされたとした場合、上述したように、必須の格助詞は「が」であるから、図7の注目統合合格フレームに記述された助詞「で」、「が」、「に」、「と」から、格助詞「が」を除く3つの助詞「で」、「に」、「と」が、オプションの助詞として認識され、さらに、そのオプションの助詞の属性として、図7の注目統合合格フレームに記述されている3つの助詞「で」、「に」、「と」それぞれの属性Instrument, Locative, Propositionが認識される。

【0177】そして、ステップS94に進み、項構造情報生成部24は、ステップS92とS93で認識した情報から、項構造情報を生成し、補助情報生成部25に出力する。

【0178】即ち、項構造情報生成部24は、例えば、図7に示した注目統合合格フレームに対応する注目動詞「目立つ」について、上述したように、必須の格助詞「が」とその属性Theme (thing/increase)のセット、並びにオプションの格助詞とその属性のセット「で」とInstrument、「に」とLocative、および「と」とPropositionが得られた場合には、図8に示した項構造情報<ArgStr:Theme (thing/increase)-(Instrument)-(Locative)-(Proposition)>を生成し、補助情報生成部25に出力する。

【0179】その後、ステップS95に進み、項構造情報生成部24が、まだ、注目統合合格フレームとしていない統合合格フレームがあるかどうかを判定する。ステップS95において、まだ、注目統合合格フレームとしていない統合合格フレームがあると判定された場合、ステップS91に戻り、項構造情報生成部24は、まだ、注目統合合格フレームとしていない統合合格フレームのうちの1つを、新たな注目統合合格フレームとし、以下、同様の処理を繰り返す。

【0180】また、ステップS95において、まだ、注目統合合格フレームとしていない統合合格フレームがないと判定された場合、下位範疇化情報生成処理を終了する。

【0181】以上のように、図1の補助情報生成装置によれば、多数のコーパスデータについて、その形態素解析結果から、基本センテンスが生成され、その基本センテンスから、不要語彙が削除される。さらに、不要語彙が削除された基本センテンスにおける動詞について、格フレームが生成され、同一の動詞についての格フレームを用いて、統合合格フレームが生成される。そして、各動詞について生成された統合合格フレームに基づいて、その動詞の下位範疇化情報と項構造情報が生成され、補助情報として出力される。従って、自然言語を構文解析や意味解析等する場合に、補助情報に含まれる下位範疇化情報や項構造情報を参照することにより、精度の高い構文解析や意味解析を行うことが可能となる。

【0182】次に、図16は、本発明を適用した自然言

語処理装置の他の一実施の形態の構成例を示している。

【0183】この自然言語処理装置は、音声によって、ユーザとの対話を行う音声対話システムを構成している。

【0184】即ち、マイク（マイクロフォン）31は、ユーザからの音声を、電気信号としての音声信号として、A/D(Analog/Digital)変換器32に供給する。A/D変換器32は、マイク31からのアナログの音声信号をA/D変換することにより、デジタルの音声データとし、音声認識部33に供給する。音声認識部33は、A/D変換器32からの音声データを、適当なフレームごとに区切り、各フレームの音声データについて音響分析を行うことにより、MFCC(Mel Frequency Cepstrum Coefficient)等の特徴ベクトルを抽出する。さらに、音声認識部33は、その特徴ベクトル系列について、例えば、HMM(Hidden Markov Model)法等によってマッチング処理を行い、マイク31に入力された音声を認識する。音声認識部33による音声の認識結果は、例えば、テキストデータで、言語処理部34に供給される。

【0185】言語処理部34は、音声認識部33からの音声認識結果を言語処理することにより、例えば、その音声認識結果に対する応答としての、例えばテキストの応答文を生成し、音声合成部35に出力する。

【0186】音声合成部35は、言語処理部34からの応答文に対応する合成音を、例えば規則音声合成処理を行うことにより生成し、D/A(Digital/Analog)変換器36に供給する。D/A変換器36は、音声合成部35からのデジタルの合成音データをD/A変換することにより、アナログの音声信号として、スピーカ37に供給する。スピーカ37は、D/A変換器36から供給される音声信号に対応する音声、即ち、言語処理部34において生成された応答文に対応する合成音を出力する。

【0187】次に、図16において、言語処理部34は、形態素解析部41、形態素解析辞書記憶部42、構文解析部43、構文解析辞書記憶部44、意味解析部45、補助情報データベース46、対話管理部47、対話履歴データベース48、および応答文生成部49から構成されている。

【0188】形態素解析部41は、音声認識部33から供給される音声認識結果について、形態素解析辞書記憶部42を参照しながら形態素解析を行い、その形態素解析結果を、構文解析部43に供給する。形態素解析辞書記憶部42は、形態素解析部41が形態素解析を行うのに参照する、例えば、形態素について、その読みや、構文属性、意味属性等が記述された形態素解析辞書を記憶している。

【0189】構文解析部43は、形態素解析部41からの形態素解析結果と、構文解析辞書記憶部44や補助情報データベース46を参照しながら、音声認識部33の

音声認識結果の構文解析を行い、その構文解析結果を、意味解析部45に供給する。構文解析辞書記憶部44は、構文解析部43が構文解析を行うに参照する、例えば、形態素の係り受け関係等についての記述がされている構文解析辞書を記憶している。

【0190】意味解析部45は、構文解析部43からの構文解析結果と、補助情報データベース46を参照しながら、音声認識部33の音声認識結果の意味解析を行い、その意味解析結果を、対話管理部47に供給する。

【0191】補助情報データベース46は、図1の補助情報生成装置としての自然言語処理装置で生成された補助情報を、多数の動詞について記憶している。

【0192】対話管理部47は、意味解析部45から供給される音声認識結果の意味解析結果や、対話履歴データベース48を参照しながら、その音声認識結果の意味内容を理解し、その音声認識結果に対応する応答文の意味内容（以下、適宜、応答内容という）を生成して、応答文生成部49に供給する。

【0193】対話履歴データベース48は、音声認識結果の意味内容や、その音声認識結果に対して、対話管理部47が生成した応答内容を、対話履歴として記憶する。

【0194】応答文生成部49は、対話管理部47からの応答内容に対応するテキストの応答文を生成し、音声合成部35に供給する。

【0195】次に、図17のフローチャートを参照して、図16の音声対話システムが行う処理（対話処理）について説明する。

【0196】マイク31に、ユーザの音声が入力され、さらに、A/D変換器32を介し、音声データが、音声認識部33に供給されると、音声認識部33は、ステップS101において、マイク31に入力された音声を音声認識し、その音声認識結果を、言語処理部34の形態素解析部41に出力して、ステップS102に進む。

【0197】ステップS102では、形態素解析部41は、音声認識部33からの音声認識結果を入力文として、その形態素解析を行い、その形態素解析結果を、構文解析部43に供給して、ステップS103に進む。ステップS103では、構文解析部43が、入力文の形態素解析結果を参照することで、その入力文に含まれる動詞についての補助情報を、補助情報データベース46から検索し、ステップS104に進む。

【0198】ステップS104では、構文解析部43が、形態素解析部41からの形態素解析結果、構文解析辞書、およびステップS103で検索した補助情報に基づき、入力文としての音声認識結果を構文解析し、その構文解析結果を、意味解析部45に供給する。さらに、ステップS104では、意味解析部45が、構文解析部43から供給される入力文としての音声認識結果の構文解析結果に基づいて意味解析を行い、ステップS105

に進む。

【0199】ステップS105では、入力文に、照応形が存在するかどうか、即ち、その入力文に含まれる動詞に必須の名詞が欠けているか（ゼロ照応形）、または必須の名詞が代名詞で代用されているかどうか判定される。

【0200】なお、入力文に、照応形が存在するかどうかは、例えば、構文解析部43による構文解析において認識することができる。

【0201】即ち、例えば、図8に示した動詞「目立つ」についての補助情報に含まれる下位範疇化情報によれば、動詞「目立つ」は、主格となる名詞句を必然的に伴うことが分かる。従って、入力文に、原形が「目立つ」の動詞が含まれている場合において、その動詞「目立つ」が、主格となる名詞句を伴っていなければ、構文解析部43は、動詞「目立つ」についての補助情報から、その動詞「目立つ」について必須の名詞句が欠けている、即ち、ゼロ照応形が存在することを認識することができる。なお、照応形の有無は、例えば、HPSG等のフレームワークにおけるサチュレーション(saturation)という機能によっても認識することができる。

【0202】ステップS105において、入力文に、照応形が存在しないと判定された場合、意味解析部45は、入力文の意味解析結果を、対話管理部47に供給し、ステップS106乃至ステップS110をスキップして、ステップS111に進む。

【0203】また、ステップS105において、入力文に、照応形が存在すると判定された場合、ステップS106に進み、意味解析部45は、補助情報データベース46を参照することにより、照応形の属性を認識する。

【0204】即ち、ステップS106では、意味解析部45は、ステップS103で検索された補助情報の下位範疇化情報と項構造情報から、入力文に含まれる動詞が必然的に伴うべき名詞の属性を認識する。そして、意味解析部45は、その入力文に含まれる動詞が必然的に伴うべき名詞の属性うち、音声認識結果に欠けている名詞、あるいは代名詞で代用されている名詞の属性を認識する。

【0205】その後、ステップS107に進み、意味管理部45は、対話管理部47に問い合わせを行うことにより、ステップS106で認識した照応形の属性と同一の属性の名詞が、対話履歴データベース48の対話履歴に存在するかどうかを判定する。

【0206】なお、ステップS107では、例えば、J. Huang, "Logical Relations in Chinese and Theory of Grammar", MIT PhD. Thesis, 1982で提唱されている、先行詞と照応家の距離はミニマルであるというヒューリスティック(Minimal Distance Principle)にしたがい、例えば、1乃至4発話前の範囲の対話履歴を対象に、照応形の属性と同一の属性の名詞が存在するかどうかを判

定する。

【0207】ステップS107において、照応形の属性と同一の属性の名詞が、対話履歴データベース48の対話履歴に存在しないと判定された場合、ステップS108に進み、対話管理部47は、ユーザに対して、照応形の内容を問い合わせる問い合わせ処理を行う。

【0208】即ち、対話管理部47は、照応形の内容を問い合わせるメッセージ（以下、適宜、問い合わせメッセージという）を、応答文生成部49に生成させ、音声合成部35およびD/A変換器36を介して、スピーカ37から、合成音で出力させる。

【0209】そして、ユーザが、問い合わせメッセージに対応して、照応形の内容を説明する発話を行うと、その音声は、マイク31、A/D変換器32、音声認識部33、形態素解析部41および構文解析部43を介して、意味解析部45に供給される。

【0210】意味解析部45は、このようにして、構文解析部43から、照応形の内容を説明するユーザの音声についての構文解析結果が供給されるのを待って、ステップS108からS109に進み、その構文解析結果に基づいて、照応形の先行詞を認識、決定して、ステップS110に進む。

【0211】一方、ステップS107において、照応形の属性と同一の属性の名詞が、対話履歴データベース48の対話履歴に存在すると判定された場合、ステップS109に進み、意味解析部43は、その対話履歴に存在する照応形と同一属性の名詞を、その照応形の先行詞として決定し、ステップS110に進む。

【0212】ステップS110では、ステップS109で決定された先行詞が、入力文の中の照応形の代わりに存在するものとして、その入力文について、構文解析部43が構文解析を行い、さらに、意味解析部45が意味解析を行い、その意味解析結果を、対話管理部47に供給する。

【0213】対話管理部47は、意味解析部45から入力文の意味解析結果を受信すると、ステップS111に進み、その意味解析結果に基づいて、入力文の意味を理解し、その入力文に対応する応答としての応答文の内容（応答内容）を生成して、ステップS112に進む。ステップS112では、対話管理部47は、入力文の意味内容と、生成した応答文の意味内容（応答内容）のセットを、対話履歴データベース48に供給して、対話履歴として記憶させるとともに、応答内容を、応答文生成部49に供給し、ステップS113に進む。

【0214】ステップS113では、応答文生成部49は、対話管理部47からの応答内容を、その意味内容とする応答文を生成し、音声合成部35に供給する。さらに、ステップS112では、音声合成部35が、応答文生成部49からの応答文に対応する合成音を生成し、D/A変換器36を介して、スピーカ37から出力させ、

対話処理を終了する。

【0215】なお、以上の対話処理においては、照応形の先行詞を、原則的には、対話履歴から決定し、対話履歴から決定することができない場合には、ユーザに問い合わせを行うようにしたが、照応形の先行詞は、対話履歴から決定し、ユーザに問い合わせを行わないようにすることも可能である。

【0216】但し、その場合には、照応形の先行詞が、同一の文の内部にあるケースと、指示や視覚を伴う理解（deictic use）が必要なものであるケースは除外する必要がある。

【0217】ここで、照応形の先行詞が、同一の文の内部にあるケースとは、照応形をproと表すと、例えば、「pro書いた論文が表彰された男」といった文が該当する。この文における照応形proは、この文で言っている男（書いた論文が表象された男）を指し示しており、照応形の先行詞となる「男」が、同一の文の内部にある。このように、照応形の先行詞が、同一の文の内部にある場合の照応形の問題は、例えば、岩波講座 言語の科学 6「生成文法」岩波書店 1997年等にあるような束縛理論(binding theory)によって解決することができる。

【0218】また、照応形の先行詞が、指示や資格を伴う理解が必要なケースとは、机の上にあるコップを指し、「それを拾え。」といった場合である。

【0219】なお、いずれのケースについても、ユーザに問い合わせを行えば、照応形の先行詞を決定することが可能である。

【0220】図17の対話処理によれば、例えば、次のようにして、照応形の先行詞が決定される。

【0221】即ち、例えば、いま、音声対話システムが、合成音「Aさんは、土用の日にうなぎを食べましたよ。」を出力し、それに対して、ユーザが、「Bさんは、もう食べたの?」と発話したとする。

【0222】この場合、音声対話システムが、ユーザの発話を正しく理解するためには、ユーザの発話「Bさんは、もう食べたの?」に、「うなぎを」を補って、「Bさんは、もう"うなぎを"食べたの?」とする必要がある。

【0223】そこで、音声対話システムは、ユーザの発話「Bさんは、もう食べたの?」に含まれる動詞（の原形）「食べる」についての補助情報を参照する。

【0224】いま、動詞「食べる」についての補助情報が、例えば、図18に示すようなものであったとする。

【0225】ここで、図18における動詞「食べる」についての補助情報の第1行目（上から1行目）は、動詞の見出し「食べる」、読み「タベル」、分類情報「他動詞」を表している。また、第2行目の下位範疇化情報<SUBCAT:NP[nom]-NP[acc]>は、動詞「食べる」が、主格(nominative)を表す名詞句(NP[nom])と対格(accurative)

を表す名詞句(NP[acc])を必然的に伴うことを表している。さらに、第3行目の項構造情報<ArgStr:Agent-Theme(food)-(Instrument)-(Locative)>は、下位範疇化情報の主格を表す名詞句NP[nom]が、動詞「食べる」の動作主(Agent)となるものであること、下位範疇化情報の対格を表す名詞句NP=[acc]が、動詞「食べる」の対象物(Theme)となるものであること、その対象物(Theme)が、食べ物(food)であること、動詞「食べる」が、必要に応じて、属性がInstrumentやLocativeで表される助詞を取り得ること、を表している。

【0226】なお、属性InstrumentとLocativeは、上述したように、それぞれ道具(例えば、「ナイフで」など)と場所(例えば、「レストランで」)を表す。

【0227】ユーザの発話「Bさんは、もう食べたの？」については、図18の補助情報を参照することにより、体格を表す名詞句であって、食べる対象物となる食べ物を表すものが欠けている(ゼロ照応形が存在する)ことが分かる。

【0228】一方、いまの場合、ユーザの発話「Bさんは、もう食べたの？」の直前に、音声対話システムが、「Aさんは、土用の日にうなぎを食べましたよ。」を出力しており、この出力のうちの「うなぎを」は、体格を表す名詞句であって、食べる対象物となる食べ物を表している。

【0229】従って、この場合、音声対話システムは、対話履歴を参照することにより、ユーザの発話「Bさんは、もう食べたの？」に欠けている、対格を表す名詞句であって、食べる対象物となる食べ物を表すものが、「うなぎ」であることを認識することができる。即ち、この場合、ユーザの発話「Bさんは、もう食べたの？」に存在するゼロ照応形の先行詞が、「うなぎ」であることが決定される。

【0230】その結果、音声対話システムは、ユーザの発話「Bさんは、もう食べたの？」に、決定した先行詞「うなぎを」を補って、「Bさんは、もううなぎを」食べたの？」とし、その意味内容を正しく理解することができる。

【0231】なお、対話履歴に、対格を表す名詞句であって、食べる対象物となる食べ物を表すものが存在しない場合には、音声対話システムは、その食べ物が何であるかを問い合わせるメッセージとして、例えば「Bさんは、何を食べたのですが？」などを生成、出力し、そのメッセージに対するユーザの返答を待って、ゼロ照応形の先行詞(いまの場合は、「うなぎ」)を決定する。

【0232】また、上述の場合には、ユーザの発話が、ゼロ照応形を有する「Bさんは、もう食べたの？」であるときを対象としたが、図17の対話処理によれば、ユーザの発話が、ゼロ照応形でない照応形を有する、例えば、「Bさんは、もう、それを(あれを)食べたの？」であるときも、ゼロ照応形における場合と同様にして、

照応形「それ(あれ)」の先行詞を決定することができる。

【0233】以上のように、図16の音声対話システムでは、動詞の下位範疇化情報と項構造情報を含む補助情報を参照することにより、入力文中に存在する照応形の属性を認識した後、その照応形の属性に基づいて、その照応形が指し示す先行詞を決定し、入力文の構文解析または意味解析を行うようにしたので、精度の高い構文解析や意味解析が可能となり、さらに、それにより、入力文の意味を正確に理解することが可能となる。

【0234】なお、本実施の形態では、補助情報に、分類情報を含めるようにしたが、補助情報は、分類情報を含めずに構成することが可能である。但し、補助情報に、明示的に、分類情報が含まれていない場合でも、下位範疇化情報から分類情報を得ることができるので、間接的には、分類情報が含まれているといえることができる。

【0235】次に、上述した一連の処理は、ハードウェアにより行うこともできるし、ソフトウェアにより行うこともできる。一連の処理をソフトウェアによって行う場合には、そのソフトウェアを構成するプログラムが、汎用のコンピュータ等にインストールされる。

【0236】そこで、図19は、上述した一連の処理を実行するプログラムがインストールされるコンピュータの一実施の形態の構成例を示している。

【0237】プログラムは、コンピュータに内蔵されている記録媒体としてのハードディスク105やROM103に予め記録しておくことができる。

【0238】あるいはまた、プログラムは、フレキシブルディスク、CD-ROM(Compact Disc Read Only Memory)、MO(Magneto Optical)ディスク、DVD(Digital Versatile Disc)、磁気ディスク、半導体メモリなどのリムーバブル記録媒体111に、一時的あるいは永続的に格納(記録)しておくことができる。このようなリムーバブル記録媒体111は、いわゆるパッケージソフトウェアとして提供することができる。

【0239】なお、プログラムは、上述したようなリムーバブル記録媒体111からコンピュータにインストールする他、ダウンロードサイトから、デジタル衛星放送用の人工衛星を介して、コンピュータに無線で転送したり、LAN(Local Area Network)、インターネットといったネットワークを介して、コンピュータに有線で転送し、コンピュータでは、そのようにして転送されてくるプログラムを、通信部108で受信し、内蔵するハードディスク105にインストールすることができる。

【0240】コンピュータは、CPU(Central Processing Unit)102を内蔵している。CPU102には、バス101を介して、入出力インタフェース110が接続されており、CPU102は、入出力インタフェース110を介して、ユーザによって、キーボードや、マウス、マイ

10

20

30

40

50

ク等で構成される入力部107が操作等されることにより指令が入力されると、それにしたがって、ROM(Read Only Memory)103に格納されているプログラムを実行する。あるいは、また、CPU102は、ハードディスク105に格納されているプログラム、衛星若しくはネットワークから転送され、通信部108で受信されてハードディスク105にインストールされたプログラム、またはドライブ109に装着されたリムーバブル記録媒体111から読み出されてハードディスク105にインストールされたプログラムを、RAM(Random Access Memory)104にロードして実行する。これにより、CPU102は、上述したフローチャートにしたがった処理、あるいは上述したブロック図の構成により行われる処理を行う。そして、CPU102は、その処理結果を、必要に応じて、例えば、入出力インタフェース110を介して、LCD(Liquid Crystal Display)やスピーカ等で構成される出力部106から出力、あるいは、通信部108から送信、さらには、ハードディスク105に記録等させる。

【0241】ここで、本明細書において、コンピュータに各種の処理を行わせるためのプログラムを記述する処理ステップは、必ずしもフローチャートとして記載された順序に沿って時系列に処理する必要はなく、並列的あるいは個別に実行される処理（例えば、並列処理あるいはオブジェクトによる処理）も含むものである。

【0242】また、プログラムは、1のコンピュータにより処理されるものであっても良いし、複数のコンピュータによって分散処理されるものであっても良い。さらに、プログラムは、遠方のコンピュータに転送されて実行されるものであっても良い。

【0243】なお、補助情報は、図16に示した音声対話システムその他、テキスト要約や翻訳その他の自然言語処理を行うシステムで用いることが可能である。また、補助情報は、図16に示したように、独立の補助情報データベース46に記憶させる他、そのシステムで用いられるレキシコン（辞書）（例えば、図17の形態素解析辞書記憶部42の形態素解析辞書や、構文解析辞書記憶部44の構文解析辞書など）に統合する形で記憶させることも可能である。

【0244】また、本発明は、日本語以外の自然言語にも適用可能である。

【0245】

【発明の効果】本発明の第1の自然言語処理装置および自然言語処理方法、並びにプログラムによれば、コーパスデータの形態素解析結果から、格フレームの生成対象とする単位である基本センテンスが生成され、その基本センテンスから、格フレームの生成に不要な語彙が削除される。さらに、不要語彙が削除された基本センテンスにおける動詞について、格フレームが生成され、同一の動詞についての格フレームに基づいて、その動詞の下位

範疇化情報と項構造情報が生成されて、補助情報として出力される。従って、その補助情報を参照することにより、精度の高い構文解析や意味解析等が可能となる。

【0246】本発明の第2の自然言語処理装置および自然言語処理方法、並びにプログラムによれば、少なくとも、動詞の下位範疇化情報と項構造情報からなる補助情報を記憶している補助情報記憶手段から、入力文に含まれる動詞についての補助情報が検索される一方、入力文中に照応形が存在するかどうかが判定され、入力文中に存在する照応形の属性が、その入力文に含まれる動詞についての補助情報に基づいて認識される。そして、照応形の属性に基づいて、照応形が指し示す先行詞が決定され、その先行詞を用いて、入力文の構文解析または意味解析が行われる。従って、精度の高い構文解析や意味解析等が可能となり、さらに、それにより、入力文の意味を正確に理解することが可能となる。

【図面の簡単な説明】

【図1】本発明を適用した自然言語処理装置の一実施の形態の構成例を示すブロック図である。

【図2】形態素解析結果を示す図である。

【図3】基本センテンスから削除される語彙（不要語彙）を説明する図である。

【図4】不要語彙が削除された形態素解析結果を示す図である。

【図5】動詞の基準形を説明する図である。

【図6】格フレームを示す図である。

【図7】統合格フレームを示す図である。

【図8】補助情報を示す図である。

【図9】補助情報生成処理を説明するフローチャートである。

【図10】基本センテンスパターン抽出処理を説明するフローチャートである。

【図11】不要語彙削除処理を説明するフローチャートである。

【図12】格フレーム生成処理を説明するフローチャートである。

【図13】動詞分類処理を説明するフローチャートである。

【図14】下位範疇化情報生成処理を説明するフローチャートである。

【図15】項構造情報生成処理を説明するフローチャートである。

【図16】本発明を適用した自然言語処理装置の他の一実施の形態の構成例を示すブロック図である。

【図17】対話処理を説明するフローチャートである。

【図18】補助情報を示す図である。

【図19】本発明を適用したコンピュータの一実施の形態の構成例を示すブロック図である。

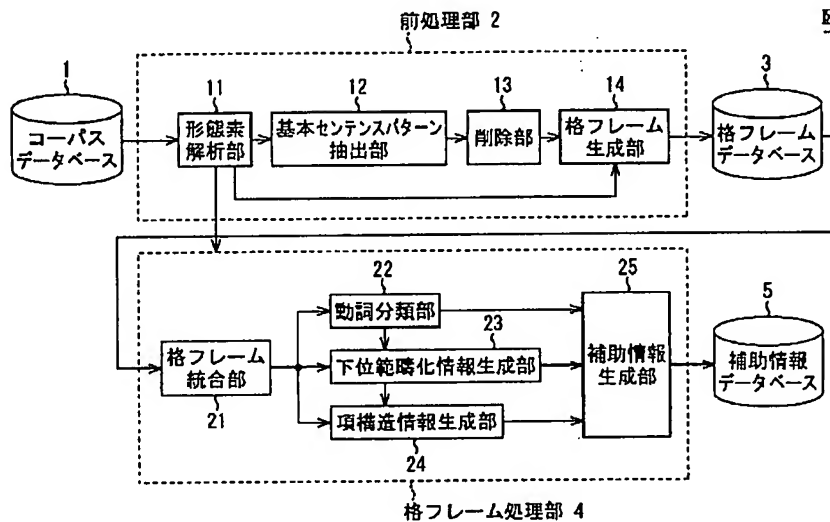
【符号の説明】

1 コーパスデータベース、 2 前処理部、 3 格

フレームデータベース, 4 格フレーム処理部, 5
 補助情報データベース, 11 形態素解析部, 1
 2 基本センテンスパターン抽出部, 13 削除部,
 14 格フレーム生成部, 21 格フレーム統合
 部, 22 動詞分類部, 23 下位範疇化情報生成
 部, 24 項構造情報生成部, 25 補助情報生成
 部, 31マイク, 32 A/D変換器, 33 音
 声認識部, 34 言語処理部, 35 音声合成部,
 36 D/A変換器, 38 スピーカ, 41 形態*

* 素解析部, 42 形態素解析辞書記憶部, 43 構
 文解析部, 44 構文解析辞書記憶部, 45 意味
 解析部, 46 補助情報データベース, 47 対話管
 理部, 48 対話履歴データベース, 49 応答文
 生成部, 101 バス, 102 CPU, 103 R
 OM, 104 RAM, 105 ハードディスク, 1
 06 出力部, 107 入力部, 108 通信部,
 109ドライブ, 110 入出力インタフェース,
 111 リムーバブル記録媒体

【図1】



自然言語処理装置(補助情報生成装置)

【図2】

| 見出し | 読み | シソーラス情報(構文属性、意味属性、動詞の原型) |
|----------------------|--------------------------------|---|
| 特に 限内果実 | トクニ ケンナイカジツ | [[CAT Adverb][VAL 特に]] [[CAT Noun][cl Compound=CN+CN][Sem food] [VAL 限内果実]] |
| が 数量 で 一八%増 | ガ スーリョー デ イチハチパーセントゾー | [[CAT Case][cl abstract][fx nominative][VAL が]] [[CAT Noun][cl CNoun][Sem amount][VAL 数量]] [[CAT Case][cl lexical][fx instrument][VAL で]] [[CAT Noun][cl Compound=Num+Classifier+suf] [Sem increase][VAL 一八%増]] |
| 金額 で 三四%増 | キンガク デ サンヨンパーセントゾー | [[CAT Punctuation][cl comma][VAL 、]] [[CAT Noun][cl CNoun][Sem money][VAL 金額]] [[CAT Case][cl lexical][fx instrument][VAL で]] [[CAT Noun][cl Compound=Num+Classifier+suf] [Sem increase][VAL 三四%増]] |
| と 伸び が 目立った | ト ノビ ガ メダッタ | [[CAT Complementizer][cl proposition][VAL と]] [[CAT Noun][cl CNoun][Sem increase][VAL 伸び]] [[CAT Case][cl abstract][fx nominative][VAL が]] [[CAT Verb][cl active][fm finite][Conj (cl2) (Stem 目立つ)(fm aff-past)(Polarity aff)(Is past)] [Style (cl plain)(fm zero)][VAL 目立った]] |
| . | . | [[CAT Punctuation][cl period][VAL 。]] |

形態素解析結果

【図3】

【図7】

図3

- (A) 副詞
[[CAT Adverb]].
- (B) 名詞+「の」(例:夏場の)
[[CAT Noun]...]
[[CAT Case][cl abstract][fx genitive][VAL の]]
- (C) 名詞+助詞+「の」(例:日本での)
[[CAT Noun]...]
[[CAT Case]...]
[[CAT Case][cl abstract][fx genitive][VAL の]]
- (D) 形容詞
[[CAT Adjective][cl stative]...]
- (E) 名詞(形容動詞語幹)+「な」(例:決定的な)
[[CAT Noun][cl AdjNoun]...]
[[CAT Verb][cl copula]...[VAL な]].
- (F) 名詞+後置詞(例:工場に対する)
[[CAT Noun]...]
[[CAT Postposition]...]
- (G) 括弧内の文書
[[CAT Punctuation][cl L-]]
[[CAT Punctuation][cl R-]]
- (H) 括弧内の文書+「の」
[[CAT Punctuation][cl L-]]
[[CAT Punctuation][cl R-]]
[[CAT Case][cl abstract][fx genitive][VAL の]]

図7

目立つ メダツ subcat: で[Instrument]:
が[increase], [thing]:
に[Locative]:
と[Proposition]:]

統合格フレーム

削除される語彙

【図4】

図4

| 見出し | 読み | シソーラス情報 |
|------|-------------|--|
| 県内果実 | ケンナイカジツ | [[CAT Noun][cl Compound=CN+CN][Sem food][VAL 県内果実]] |
| が | ガ | [[CAT Case][cl abstract][fx nominative][VAL が]] |
| 数量 | スーリョー | [[CAT Noun][cl CNoun][Sem amount][VAL 数量]] |
| で | デ | [[CAT Case][cl lexical][fx instrument][VAL で]] |
| 一八%増 | イチハチパーセントゾー | [[CAT Noun][cl Compound=Num+Classifier+suf][Sem increase][VAL 一八%増]] |
| 、 | 、 | [[CAT Punctuation][cl comma][VAL 、]] |
| 金額 | キンガク | [[CAT Noun][cl CNoun][Sem money][VAL 金額]] |
| で | デ | [[CAT Case][cl lexical][fx instrument][VAL で]] |
| 三四%増 | サンヨンパーセントゾー | [[CAT Noun][cl Compound=Num+Classifier+suf][Sem increase][VAL 三四%増]] |
| と | ト | [[CAT Complementizer][cl proposition][VAL と]] |
| 伸び | ノビ | [[CAT Noun][cl CNoun][Sem increase][VAL 伸び]] |
| か | カ | [[CAT Case][cl abstract][fx nominative][VAL が]] |
| 目立った | メダッタ | [[CAT Verb][cl active][fm finite][Conj](cl2)(Stem 目立つ)(fm aff-past)(Polarity aff)(Ts past)][Style(cl plain)(fm zero)][VAL 目立った]] |

削除処理後の基本センテンス

【図5】

図5

| 見出し | 読み | シソーラス情報 |
|------|------|---|
| 目立つ | メダツ | {[CAT Verb][cl active][fm finite][Conj(c)2] (fm aff-non-past)(Stem 目立つ) (Polarity aff)(Ts non-past)}... |
| 目立った | メダッタ | {[CAT Verb][cl active][fm finite][Conj(c)2] (fm aff-past)(Stem 目立つ) (Polarity aff)(Ts past)}... |

(A)

| 見出し | 読み | シソーラス情報 |
|------|--------|---|
| 適用する | テキヨースル | {[CAT Noun][cl Vnoun]...[VAL 適用]} {[CAT Verb][cl active][fm finite]...[Stem する] (fm aff-non-past)...[VAL する]} |

(B)

| 見出し | 読み | シソーラス情報 |
|------|------|--|
| 見込んで | ミコンデ | {[CAT Verb]...[fm infinite]...[Stem 見込む] (fm pres-participle)...[VAL 見込んで]} |
| いる | イル | {[CAT Verb][cl active][fm finite]...[Stem いる] ...[VAL いる]} |

(C)

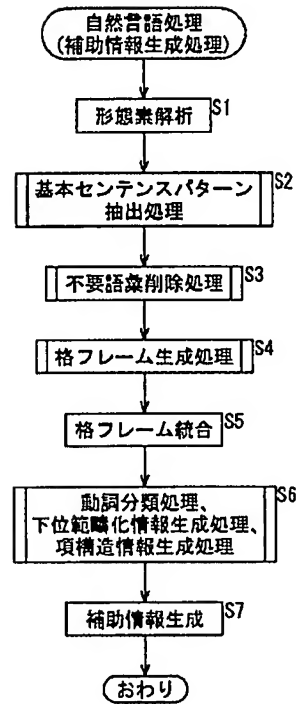
| 見出し | 読み | シソーラス情報 |
|------|--------|--|
| 展開して | テンカイシテ | {[CAT Noun][cl Vnoun]...[VAL 展開]} {[CAT Verb]...[fm finite]...[Stem する] (fm pres-participle)...[VAL して]} |
| いる | イル | {[CAT Verb]...[fm finite]...[Stem いる]... [VAL いる]} |

(D)

動詞の基準形

【図9】

図9



【図6】

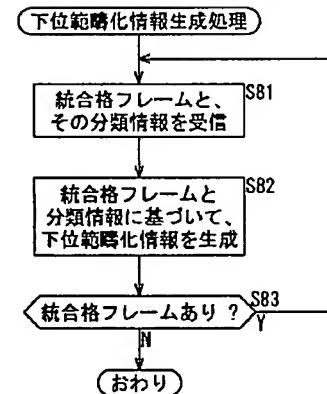
[目立つ C_FRAME: で[instrument], が[increase]]
 [目立つ C_FRAME: が[thing]]
 [目立つ C_FRAME: と[proposition], が[thing]]
 [目立つ C_FRAME: で[instrument], に[locative], が[increase]]

格フレーム

【図14】

図14

図14



【図8】

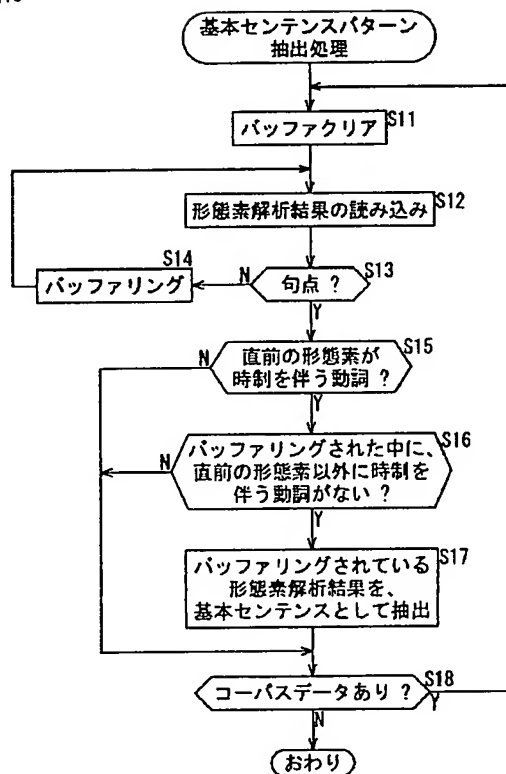
8

[目立つ メダツ 能動動詞
 下位範疇化情報: <SUBCAT:NP[nom]>
 項構造情報: <ArgStr:Theme[thing/increase]
 -(Instrument)-(Locative)-(Proposition)>
]

補助情報

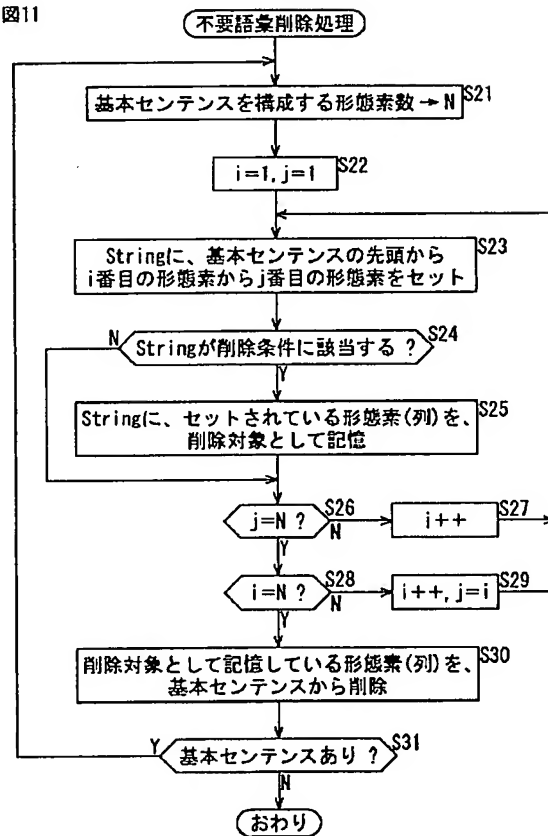
【図10】

図10



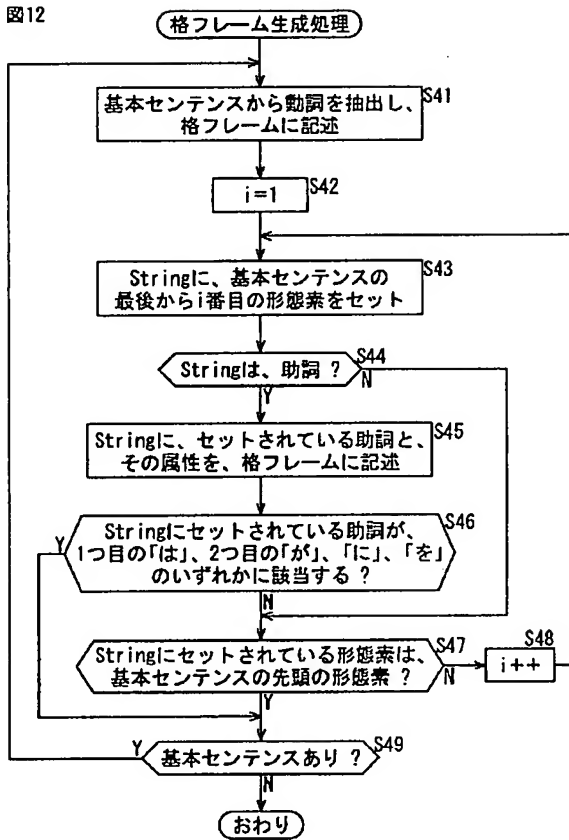
【図11】

図11



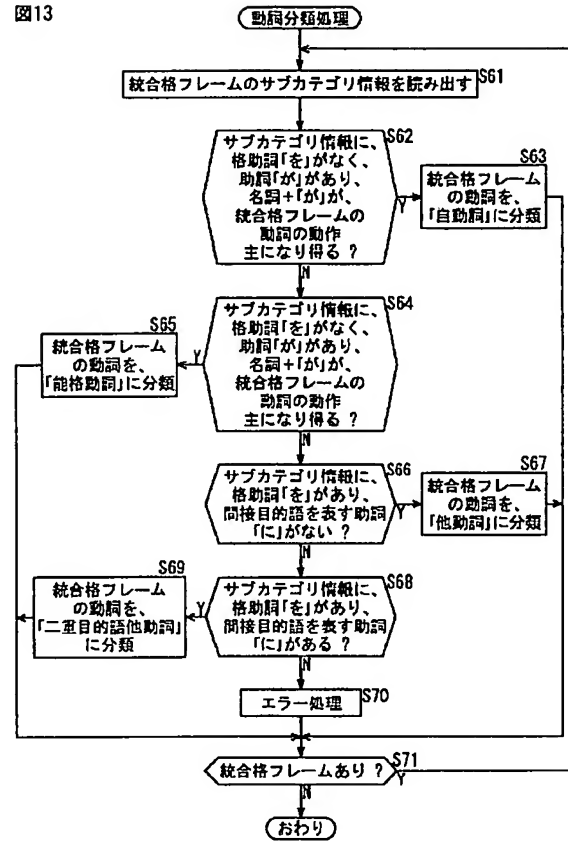
【図12】

図12



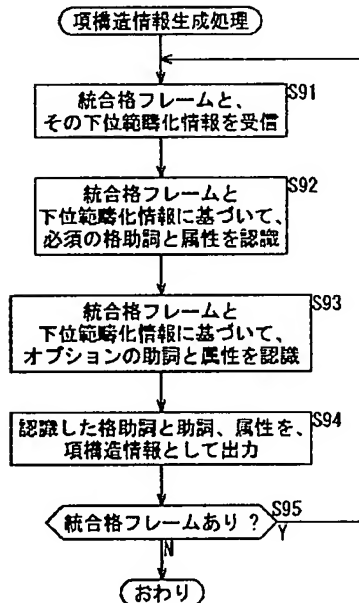
【図13】

図13

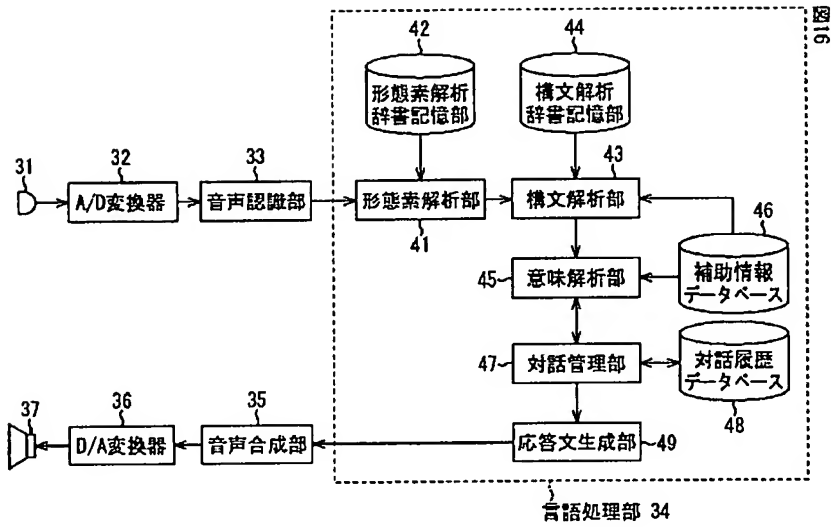


【図15】

図15



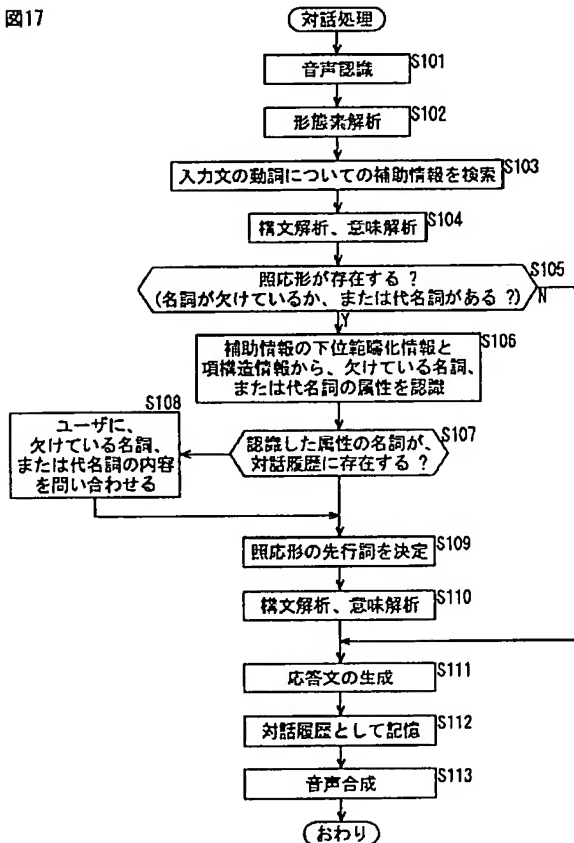
【図16】



自然言語処理装置(音声対話システム)

【図17】

図17



【图 18】

图 18

食べる タベル

他動詞

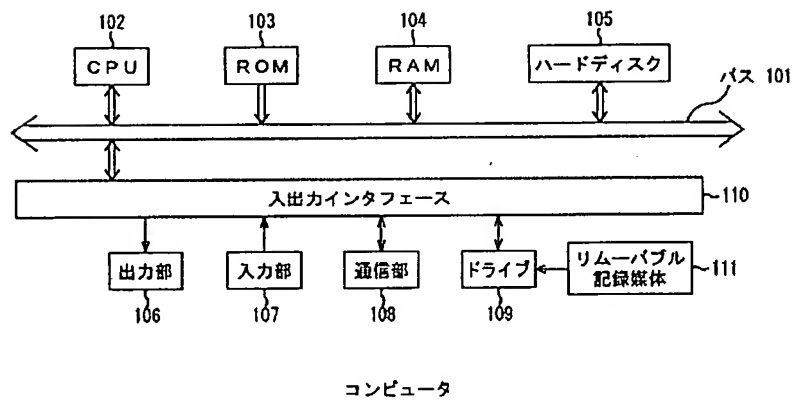
下位範疇化情報: (SUBCAT: NP[nom]-NP[acc])
項構造情報: (ArgStr: Agent-Theme {food}
-(Instrument)-(Locative))

}

「食べる」についての補助情報

【図 19】

图 19



フロントページの続き

(72)発明者 下村 秀樹
東京都品川区北品川6丁目7番35号 ソニ
ー株式会社内

Fターム(参考) 5B091 AA15 AB15 AB19 CA02 CA12
CA14 CC01 CC15